THE LIBRARY

UNIVERSITY OF MAINE
ORONO, MAINE

STATISTICAL METHODS

APPLIED TO ECONOMICS AND BUSINESS

BY

FREDERICK CECIL MILLS

ASSOCIATE PROFESSOR OF BUSINESS STATISTICS
COLUMBIA UNIVERSITY

NEW YORK
HENRY HOLT AND COMPANY

COPYRIGHT, 1924,
BY
HENRY HOLT AND COMPANY

PRINTED IN THE
UNITED STATES OF AMERICA

To
D. C. M.

76.751

PREFACE

The last decade has witnessed a remarkable stimulation of interest in quantitative methods in business and in the social sciences. The day when intuition was the chief basis of business judgment and unsupported hypothesis the mode in social studies seems to have passed. Following the lead of workers in the older and traditionally more accurate physical sciences, social scientists and serious students of business are employing in greater measure than ever before a method of study based upon the observation and analysis of *facts*. When these observations are quantitative in character appropriate methods are necessary for their organization and interpretation. This book deals with methods of combining and analyzing such observations, with primary emphasis upon materials drawn from the fields of economics and business.

The justification for limiting the treatment to these particular fields is twofold. Although general statistical methods are practically universal in their application, special problems are encountered in every field of study. This is particularly true in the realm of economics, which presents many distinctive difficulties and many characteristic problems. Methods which are in some degree specialized to meet these particular requirements have been developed, and these methods call for treatment in a work which is restricted in scope. In the second place, methods can be most effectively explained in terms of particular subjects; abstract methodology is barren of interest to the average person. For these reasons the book has been written with reference to the specific needs of quantitative workers in economics and business.

In the explanation of methods no attempt has been made to secure the brevity of exposition which may be desirable in a strictly mathematical work. The purpose throughout has been to write for the learner not for the finished master, and the explanations have been prepared with the needs of the former in mind. I have felt free to omit certain detailed demonstrations of theorems

because this book is presented as an introduction to the subject, not as an exhaustive treatise.

The methods of quantitative analysis which are in general use today represent a long accretion, an accumulation of contributions from workers in many fields. It would be vain to attempt to enumerate all the individuals who have contributed to the development of the science of statistics. Individual references are given in particular cases in the body of the text, but no list of such acknowledgments can serve as a complete record of the debt modern statisticians owe to their predecessors.

For assistance in the preparation of the material contained in this book I am under many obligations. To Mr. H. E. Anderson and Mr. H. B. Killough I am indebted for certain of the data employed in Chapters XI, XIII and XIV. Professor Warren M. Persons of the Harvard Committee on Economic Research has courteously permitted me to make use of certain results of his work on commodity price index numbers. The index of building costs used in Chapter VII and the index of business conditions presented in Chapter IX are products of the Statistical Division of the American Telephone and Telegraph Company. I have employed them with the kind permission of Mr. Seymour L. Andrew, Chief Statistician. Suggestions from Professor A. H. Mowbray of the University of California have enabled me to remove several obscurities which were present in an earlier mimeographed edition. I am deeply grateful to Professors Henry L. Moore and Theodore H. Brown of Columbia University and to Mr. Henry Schultz of the Institute of Economics for their help in critically reviewing portions of the manuscript. Mr. Herbert F. Tutt has rendered valuable aid in the preparation of the charts. For assistance at every stage of the work involved in the writing of this book I am under deep obligation to Mr. Donald H. Davenport. His aid in the collection of material, in the preparation of charts and in the onerous task of seeing the book through the press has been invaluable. To my wife, above all others, I am indebted for a measure of constant and generous help which cannot be adequately acknowledged here.

F. C. M.

November, 1924.

CONTENTS

CHAPTER		PAGE
I.	Statistical Methods and the Problems of Economics and Business	3
II.	Graphic Presentation	11
III.	The Organization of Statistical Data: The Frequency Distribution	61
IV.	Description of the Frequency Distribution: Averages	97
V.	Description of the Frequency Distribution: Measures of Variation and Skewness	147
VI.	Index Numbers of Prices	169
VII.	The Analysis of Time Series: Measurement of Trend	252
VIII.	The Analysis of Time Series: Measurement of Seasonal and Cyclical Fluctuations	315
IX.	Index Numbers of Physical Volume	344
X.	The Measurement of Relationship: Linear Correlation	363
XI.	The Measurement of Relationship between Time Series	410
XII.	The Measurement of Relationship: Non-linear Correlation	433
XIII.	The Measurement of Relationship and the Problem of Estimation	453
XIV.	The Measurement of Relationship: Multiple and Partial Correlation	485
XV.	Elementary Probabilities and the Normal Curve of Error	516
XVI.	Statistical Induction and the Problem of Sampling	548
Appendix A.	The Method of Least Squares as Applied to Certain Statistical Problems	562
Appendix B.	Glossary of Symbols	582
List of References		589
Index		595

LIST OF CHARTS

FIGURE		PAGE
1.	Location of a Point with Reference to Rectangular Coordinates	12
2.	Cotton Prices in the United States, by Months, During the Year 1923	15
3.	Graph of the Equation $y = x$	17
4.	Graph of the Equation $y = 3x + 2$	19
5.	Parabola: Graph of the Equation $y = x^2$	21
6.	Equilateral Hyperbola: Graph of the Equation $y = x^{-1}$	22
7.	Exponential Curve: Graph of the Equation $y = 2^x$	24
8.	Sine Curve: Graph of the Equation $y = \text{Sin } x$	26
9.	Graph of the Equation $\log y = 2 \log x$	31
10.	Graph of the Equation $y = x^2$. (Plotted on paper with logarithmic scales)	33
11.	The Compound Interest Law: The Growth of $10.00 at Compound Interest at 6% for 100 Years. (Plotted on the arithmetic scale)	34
12.	The Compound Interest Law: The Growth of $10.00 at Compound Interest at 6% for 100 Years. (Plotted on the semi-logarithmic or ratio scale)	35
13.	Wheat Flour Exports from the United States, 1901–1923	38
14.	Production of Steel in the United States, 1896–1922. (Plotted on the semi-logarithmic scale)	39
15.	Sales of the Acme Corporation, 1910–1923. Showing the Total Sales in the United States and the Sales in Certain Subdivisions. (Plotted on the arithmetic scale)	40
16.	Sales of the Acme Corporation, 1910–1923. Showing the Total Sales in the United States and the Sales in Certain Subdivisions. With Scales of Increase, Decrease and Comparison. (Plotted on semi-logarithmic paper)	41
17.	Farms in New England States in 1920	45
18.	Elements of Production Costs of the XYZ Corporation, by Months, 1923	47

FIGURE PAGE

19. Comparison of Scheduled and Actual Output (Cumulative) Speedwell Automobile Company, 1924 48
20. Comparison of Scheduled and Actual Output: Gantt Progress Chart. (Showing the situation on February 28th) 50
21. Comparison of Scheduled and Actual Output: Gantt Progress Chart. (Showing the situation on September 30th) 50
22. Column Diagram: Distribution of 210 Employees Classified on the Basis of Weekly Earnings. (Class-interval = $2.00) 74
23. Column Diagram: Distribution of 210 Employees Classified on the Basis of Weekly Earnings. (Class-interval = $1.00) 75
24. Column Diagram: Distribution of 210 Employees Classified on the Basis of Weekly Earnings. (Class-interval = $.50) 76
25. Column Diagram: Distribution of 210 Employees Classified on the Basis of Weekly Earnings. (Class-interval = $.25) 77
26. Frequency Polygon: Distribution of 210 Employees Classified on the Basis of Weekly Earnings. (Class-interval = $2.00) 78
27. Frequency Polygon: Distribution of 210 Employees Classified on the Basis of Weekly Earnings. (Class-interval = $1.00) 78
28. Frequency Polygon: Distribution of 210 Employees Classified on the Basis of Weekly Earnings. (Class-interval = $.50) 79
29. Column Diagram: Distribution of Personal Income Recipients in the United States, 1918. Including all Recipients of Incomes below $4,000. (Class-interval = $500) 83
30. Column Diagram: Distribution of Personal Income Recipients in the United States, 1918. Including all Recipients of Incomes below $4,000. (Class-interval = $200) 84
31. Column Diagram: Distribution of Personal Income Recipients in the United States, 1918. Including all Re-

FIGURE PAGE

cipients of Incomes below $4,000. (Class-interval = $100) 85

32. Frequency Curve: Distribution of Personal Income Recipients in the United States, 1918. Including all Recipients of Incomes below $4,000. (Derived from the column diagram with class-interval of $100) 86

33. Cumulative Frequency Curve: Distribution of Telephone Poles Classified according to Length of Life. (Cumulated upward) 91

34. Cumulative Frequency Curve: Distribution of Telephone Poles Classified according to Length of Life. (Cumulated downward) 92

35. Distribution of Sawmills in the United States Classified according to Labor Cost in 1920. Illustrating the Structural Relation between the Ogive and the Frequency Curve 94

36. Frequency Curve: Distribution of 18,780 Soldiers Classified according to Height 98

37. Frequency Curve: Distribution of Errors of Observation 99

38. Zone of Dispersion, Artillery Firing, Showing the Theoretical Percentage Distribution of Shots 100

39. Column Diagram: Distribution of 1,000 Shots from a Single Gun 102

40. Frequency Polygon: Distribution of Heads in a Coin-Tossing Experiment 103

41. Frequency Polygon: Distribution of 5,540 Cases of Change in the Wholesale Prices of Commodities from One Year to the Next 104

42. Frequency Polygon: Distribution of London-New York Exchange Rates 105

43. The Normal Curve of Error 108

44. Illustrating the Location of the Median with Ungrouped Data (Personal Incomes of Seven Individuals) . . 121

45. Distribution of Weekly Earnings of Employees. A Smoothed Frequency Curve, Showing the Relation between Mean, Median and Mode 132

46. Cumulative Distribution of Weekly Earnings of Employees, Illustrating the Graphic Location of Median and Quartiles 134

FIGURE PAGE

47. Frequency Polygon: Distribution of Relative Prices of 346 Commodities in 1914. (Average prices in 1913 = 100) 180
48. Frequency Polygon: Distribution of Relative Prices of 222 Commodities in 1900. (Average prices in 1890 = 100) 183
49. Frequency Polygon: Distribution of Relative Prices of 1,437 Commodities in 1918. (Average prices July, 1913 to June, 1914 = 100) 186
50. Frequency Polygon: Distribution of Relative Prices of 1,437 Commodities in 1918, with Relative Prices Plotted on Logarithmic Scale. (Average prices July, 1913 to June, 1914 = 100) 189
51. Comparison of Five Simple Index Numbers of Farm Crop Prices, 1910–1923. (1910 = 100) 202
52. Comparison of Four Weighted Index Numbers of Farm Crop Prices, 1910–1923. (1910 = 100) 217
53. A Graphic Representation of the Relation between Certain Elements of the Price System 224
54. New York Bank Clearings, 1860–1923, with Moving Averages 260
55. Illustrating the Fitting of a Straight Line to Nine Points 273
56. Illustrating the Fitting of a Second Degree Parabola to Nine Points 283
57. Business Failures in the United States, 1897–1921, with Three Lines of Trend 289
58. Production of Crude Petroleum in the United States, 1908–1922, with Line of Trend Fitted to Logarithms 293
59. Production of Crude Petroleum in the United States, 1908–1922, with Line of Trend Fitted to Natural Numbers 295
60. Comparison of Actual and Deflated Values of Building Contracts Awarded, by Months, 1914–1923 313
61. Frequency Distributions of Ratios of Monthly Egg Prices to the Corresponding Ordinates of Secular Trend 325
62. Illustrating the Method of Determining Monthly Trend Values 330
63. Showing the Relation between Trend Values of Total

FIGURE PAGE

Annual Production, Average Monthly Production and Production by Months 332
64. Production of Bituminous Coal in the United States, 1913–1923 (by months) 339
65. Cyclical and Accidental Fluctuations in Bituminous Coal Production in the United States, 1913–1923 . . 341
66. Composite Index of General Business Activity 356
67. Scatter Diagram showing the Relation between Taxable Personal Incomes and Motor Vehicle Registration, by States, in 1921, with Line of Average Relationship . 366
68. Tabulation of Items in a Correlation Table 380
69. Scatter Diagram of Federal Reserve and Commercial Bank Rates, with Line of Average Relationship and Zones of Estimate 385
70. Showing the Relation between Discount Rates of Commercial Banks and Federal Reserve Bank Discount Rates. (Commercial bank rates dependent) 395
71. Showing the Relation between Federal Reserve Bank Discount Rates and the Discount Rates of Commercial Banks. (Federal Reserve bank rates dependent) 397
72. Showing the Relation between Number of Wage Earners in Factories and Value of Products in Ten Selected Cities in the State of New York 407
73. Showing the Relation between Number of Wage Earners in Factories and Value of Products in Eleven Selected Cities in the State of New York 408
74. Cotton Production in the United States, Crop Years 1900–1901 to 1922–1923, with Lines of Trend . . . 413
75. Prices of Middling Upland Cotton in New York, Crop Years 1900–1901 to 1922–1923, with Lines of Trend 415
76. Comparison of the Cyclical Fluctuations in Industrial Stock Prices and in General Business Activity, 1903–1923 423
77. Coefficients of Correlation between Index of Industrial Stock Prices and Index of Business Activity, 1903–1914, showing the Results Secured with Different Pairings 424
78. Coefficients of Correlation between Index of Industrial

FIGURE PAGE

Stock Prices and Index of Business Activity, 1919–1923, showing the Results Secured with Different Pairings 425

79. Scatter Diagram showing the Relation between Alfalfa Yield and Irrigation Water Applied, with Two Lines of Regression 433

80. Scatter Diagram showing the Relation between Wheat Yield and Nitrogen Applied as Fertilizer, with Straight Line of Regression and Line Joining the Means of the Columns 444

81. The Relation between the Production and Price of Oats; Illustrating the Use of an Arithmetic Equation of Regression and Arithmetic Zones of Estimate . . . 479

82. The Relation between the Production and Price of Oats; Illustrating the Use of a Logarithmic Equation of Regression and Geometric Zones of Estimate . . . 480

83. The Relation between the Production and Price of Oats; Illustrating the Use of a Logarithmic Equation of Regression and Geometric Zones of Estimate. (Plotted on double logarithmic paper) 481

84. The Relation between the Production and Price of Oats; Illustrating the Use of an Equation of Regression Based upon Reciprocals and of Harmonic Zones of Estimate 483

85. A Comparison of Actual and Theoretical Frequencies in a Dice-rolling Experiment 525

86. Illustrating the Fitting of a Normal Curve to Frequency Distribution of Telephone Subscribers, Classified According to Message Use 536

87. An Illustration of the Measurement of Areas Under the Normal Curve 539

STATISTICAL METHODS

CHAPTER I

STATISTICAL METHODS AND THE PROBLEMS OF ECONOMICS AND BUSINESS

The term *business* as commonly used today covers many kinds of human activity. Since we are concerned with the development of methods of handling economic data and solving business problems it is appropriate that the particular types of business activity in connection with which statistical methods may be employed be briefly described.

Classes of Business Activity

The tasks which confront business men may, without undue straining, be placed in three classes. Many business men are concerned with but one type of activity, it is true, yet the world of business must deal with problems arising in all three of the fields to be outlined. First, in logical sequence, are the technical tasks which arise in the processes of production, involving problems of chemistry and physics, of engineering, of animal husbandry, of navigation. The basic technical knowledge called for in the solution of these problems furnishes the foundation of our economic life. This is the domain of the hard-won arts of handling the raw materials and controlling the forces of nature.

In the second class come those activities which are connected with the internal organization and administration of individual business units. The technical functions of manipulating organic and inorganic matter for the satisfaction of human wants are performed through administrative units, single farms, mines, factories, railroads, department stores. A whole new division of problems is faced by the

business man in organizing these units, in coördinating the work of the different departments, in supervising the daily activities of the individuals making up each organization. While these are perhaps less fundamental than the technical problems of production, they are, for the average business man, more pressing and more difficult. Scientific method has made less progress in solving these latter problems. There is not the organized body of knowledge which is found in the former field, nor are there the same trained experts to whom the tasks may be delegated.

The two types of economic activity which have been named above include tasks which are in a sense self-centered and controllable. The manufacturer of steel has his technical problems of smelting and refining, his particular administrative duties. The farmer or mine-owner faces the same types of problems, in forms peculiar to his own situation. In the performance of tasks in these fields each man is dealing with problems all the elements of which are under more or less perfect control. Difficulties arise, but these are ordinarily difficulties inherent in the given task, not difficulties arising from a sudden change in the constituent elements of the problem, or the sudden interjection of a new factor. In this respect the third category of tasks to be performed by the business man differs materially from the first two. For this class is composed of problems the most characteristic feature of which is that the elements are not subject to control by the individuals directly concerned.

This third division includes buying and selling, and all the attendant activities which are carried on in terms of prices. As economic life is at present organized these functions are, to the business man, the most important ones he performs. The technical tasks of production and of internal organization and administration are but means to an end. For the business man the goal of economic activity is the disposal of his product at a profit. The

tasks preliminary to this final sale are of necessity subordinated to it, and so performed that the final aim may be achieved. The point of emphasis here is that the business man, in buying and selling, faces problems containing elements which he cannot control. In securing his raw material, in bringing together the other agents needed in production, and in the final disposal of his product, the business man deals with markets — commodity markets, labor markets, money markets — and finds himself acting in relation to a system of prices quite beyond his control in its major movements. The other less fundamental phases of his activity are subject to a high degree of control, but when the business man comes to the final and most important act, the profitable sale of his product, his power of control dwindles. The motivating force in business activity is the hope of pecuniary profits, pecuniary profits depend upon successful buying and selling, successful buying and selling depend upon favorable conditions in an uncontrollable world of prices — here is the argument which states the major problem of business. And these are the facts which make the price system the dominating and all-important factor in modern business life.

The modern entrepreneur lives in an environment of prices. The term "environment" is not an unapt figure; this world of prices in which the business man functions constitutes a coherent, consistent, well-articulated system of interdependent parts, a system which encompasses all the business activities of the entrepreneur. Since the system is beyond the control of the individual he must adapt himself to it, and must base his activities upon as complete an understanding of the system as he may obtain. Without this understanding the major problems of business are incapable of solution.

Quantitative Character of Economic and Business Problems

Problems falling in the first of the classes outlined above have long been recognized as essentially quantitative in character. Their solution calls for the application of the methods of precision which have been developed in the physical sciences. It is no less true that the strictly economic and business problems which fall in the other classes require the employment of quantitative methods. Qualitative, non-measurable considerations may enter in the solution of such problems, but these factors generally rest upon a quantitative basis. Facts, measured, weighed and compared with other facts, constitute the basis of business judgments and the foundation of economic reasoning. Statistical methods are merely methods of measuring, weighing and comparing facts.

Of the three classes of problems distinguished in the preceding section two come within the scope of the present discussion. Though the methods of statistics are in part applicable to the solution of technical problems of production, it is not the purpose of the present work to develop this subject. For the solution of problems in the two other fields — those connected with the internal organization and administration of business units and with the processes of buying and selling which bring the business man into contact with the price system — methods of statistical analysis are peculiarly appropriate.

Statistical Methods and Problems of Internal Administration

The typical business man, in the administration of his organization, is called upon to deal with masses of quantitative data, quantitative in the sense that they are based

upon measurements in terms of specific units. He is dealing with tons of coal, cubic feet of gas, or kilowatt hours of energy consumed; with tons of pig iron or pairs of shoes produced; with machine hours and man hours; with wages, costs of production and selling prices expressed in dollars and cents. With the increasing size of the business unit the data with which the administrator must deal become increasingly complicated and numerous, and it becomes increasingly difficult to determine their true significance. Under intuitive or rule-of-thumb methods of administration it is impossible effectively to analyze large masses of data and to control business units above the average in size. It has been abundantly demonstrated that the law of decreasing returns comes into play in business largely because of administrative difficulties.

Whenever one deals with masses of data the problem is one of condensation and analysis — condensation and simplification in order that it may be possible for limited human faculties to handle the data, analysis (and comparison) in order that the elements in the problem may be distinguished and their significance appreciated. Statistical methods have been developed to facilitate the condensation and analysis of masses of quantitative data.

As a typical example of such a problem may be mentioned the allocation of costs, an operation which has been called cost accounting. The proper analysis of all the factors which enter into this problem is only possible through the use of statistical methods. Accounting methods, necessarily restricted to the treatment of pecuniary units, are inadequate for the complete analysis of the items of expense. Quite apart, too, from the knowledge to be derived from an allocation of costs, expense items should be analyzed statistically, variations noted, comparisons made, and wastes discovered. The analysis of sales records, again, calls for the condensation of masses of data, their representation in simple, understandable form, and the

determination of the significance of the data. The analysis of markets, the study of purchasing records and of commodities, require the use of quantitative methods not restricted in their application to any one class of measurements. At every hand in internal administration statistical methods may be used to supplement accounting methods, to extend the knowledge of the executive, and to make more effective the control of business operations.

Statistical Methods and External Problems

New problems are encountered when the business man goes into the market to buy or sell. The price system, the movements of which are of such fundamental interest to the business man, requires analysis through the use of quantitative methods. So complex and numerous are the data to be dealt with here that simplification is imperative. Again, he is faced with the phenomena of the business cycle, and if he is to adapt his business policies to the swings of the cycle he must undertake the analysis of these phenomena, employing tools appropriate to the task. Apart somewhat from the immediate interests of the business man, but of great interest to the economist, are all the problems connected with the economic process of distribution, the allocation of income and wealth among the agents of production. These, as well as that other great economic problem concerned with the question of value or price determination, are quantitative problems, to be solved through the use of quantitative methods of research.

What are these methods, and wherein does research employing such methods differ from other types of research? Scientific inquiry, whatever its particular method may be, proceeds through careful observation, logical inference and accurate verification. Quantitative methods differ from others only in that observation, inference and verification are based upon *measurement*, and are therefore more likely

to be exact and accurate than are non-quantitative methods of analysis. Until measurement is possible in a science it is unavoidable that its observations and findings should lack precision, no matter how brilliant the flashes of intuition nor how painstaking the labors of its students may be. The employment of methods of measurement, making possible the analysis of the factors involved in terms of precise units, gives to a science all the advantages which sharp-edged tools have over blunt and unreliable instruments. Mathematics and its offspring, statistics and accounting, are the powerful instruments which the modern economist has at his disposal, and of which business, through the development of research agencies and methods, is making constantly greater use.

The tools of the statistician are merely certain mathematical methods, developed for particular types of research. These types of research were not economic in the original development of statistical methods, but social, political and anthropometric, with one line of development (that relating to the theory of probabilities) extending back through the field of logic to the gaming table. Yet these tools, developed for work in restricted spheres, have been found to possess much wider applicability, and economics has been one of the newer fields in which the application of these methods, with appropriate alterations and additions, has had fruitful results. The economist has found his hand strengthened and the precision of his work materially increased by the new tools. And business, together with the more abstruse science of economics, has profited.

It is perhaps unnecessary at this point to stress the limitations of the statistical method. For limitations it has, and limitations not always kept in view by those employing it. A discussion of this subject may be deferred to a later point. The fact may be emphasized at this stage, however, that the statistical method, as a tool, requires intelligent usage, and that the results secured through statisti-

cal analysis require intelligent interpretation. Man has not yet invented a machine so automatically perfect that jumbled facts may be fed in at one end, while answers to problems flow out at the other. The faculties of intelligent reasoning and critical judgment have not become out-worn appendages because of the development of statistical methods of research.

CHAPTER II

GRAPHIC PRESENTATION

The explanation of methods of condensing, analyzing and interpreting the facts of business and economics must start with the discussion of some fundamental considerations which are mathematical rather than statistical in character. In doing so it is deemed advisable, even at the risk of treading quite familiar ground, to explain certain mathematical conceptions to which constant reference will be made in later chapters.

Statistical analysis is concerned primarily with data based upon measurement, expressed either in pecuniary or physical units. The methods of coördinate geometry, developed first by the philosopher Descartes, greatly facilitate the manipulation and interpretation of such data. A summary of the basic principles of coördinate geometry will not be out of place.

Rectangular Coördinates

If two straight lines intersecting each other at right angles are drawn in a plane, it is possible to describe the location of any point in that plane with reference to the point of intersection of the two lines. We will call one of the lines (a vertical line) $Y'Y$, the other line (horizontal) $X'X$, and the point of intersection (or *origin*) O (cf. Fig. 1). If P be any point in the plane, we may draw the line PM, parallel to $Y'Y$ and intersecting $X'X$ at M, and the line PN, parallel to $X'X$ and intersecting $Y'Y$ at N. If we set OM equal to a units and ON equal to b units, a and b constitute the *coördinates* of P, describing its location with respect to the origin, O. Thus, in the figure, a equals 6 and

b equals 5. The distance a along the x-axis is termed the *abscissa* of the point P, while the distance b along the y-axis is termed the *ordinate* of the point P. (It is a rule of notation always to give the abscissa first, followed by the ordinate.) The coördinates of any other point in the same plane may be determined in the same way. Conversely,

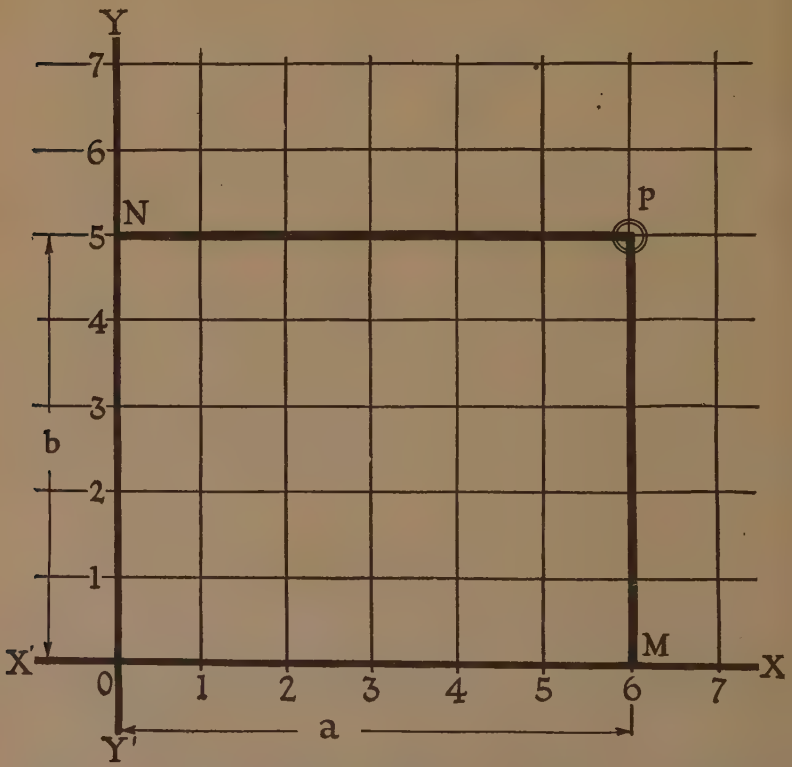


Fig. 1. — Location of a Point with Reference to Rectangular Coördinates

any two real numbers determine a point in the plane, if one be taken as the abscissa and the other as the ordinate.

A point may lie either to the right or left or above or below the origin, O. It is conventional to designate as positive abscissas laid off to the right of the origin, and as negative abscissas laid off to the left of the origin, while

ordinates are positive when laid off above the origin and negative when laid off below the origin. In general, the values to be dealt with in economic statistics lie in the upper right-hand quadrant, where both abscissa and ordinate are positive.

This conception of coördinates is fundamental in mathematics and of basic importance in statistical work. A very simple example will illustrate the utility of this device in representing business data. The figures presented in the following table may be employed.

TABLE 1

Price of Cotton, by Months, During the Year 1923

(Average prices received by producers in the United States)

Month	*Price per pound*
January	24.5 cents
February	25.9 "
March	27.7 "
April	28.4 "
May	26.9 "
June	25.6 "
July	26.2 "
August	23.5 "
September	24.1 "
October	27.2 "
November	28.8 "
December	31.0 "

(From *Weather, Crops and Markets*, Dec. 29, 1923.)

These data may be represented graphically on the coordinate system, months being laid off along the x-axis and prices along the y-axis, as in the accompanying diagram (Fig. 2). In plotting the abscissas, January, 1923, is considered as located at the point of origin. The x-value of the entry for January, 1923, is thus 0, of the February figure 1, etc. The coördinates of the point representing the price of cotton in January, 1923, are 0, 24.5; for February the values are 1, 25.9. The coördinates for December are 11, 31.0. The trend of cotton prices during

the year may be more easily followed if the points are connected by a series of straight lines, as is done in the figure.

Independent and Dependent Variables

In the location of any point by means of coördinates it has been pointed out that two values are involved; every point ties together and expresses a relation between two factors. In the above case these are months and cotton prices. With the passage of time the price of cotton changes, and the broken line shows the direction and magnitude of these changes. Both time and price are *variables*, that is, they are quantities not of constant value but characterized by variations in value in the given discussion. Thus in Fig. 1 the abscissa has a fixed value of 6, while the ordinate has a fixed value of 5, but in Fig. 2 both abscissa and ordinate have varying values, the one varying from 0 to 11, the other from 23.5 to 31.0. The symbols x and y are, by convention, used to designate such variable quantities as these, the former in all cases representing the variable plotted along the horizontal axis, the latter representing the variable plotted along the vertical axis.[1]

In Fig. 2, which depicts the changes taking place in cotton prices with the passage of time, it will be noted that the latter variable changes each time by an arbitrary unit, one month. Having made an independent change in the time factor we then determine the change in price taking place during the period thus arbitrarily chopped out. The variable which increases or decreases by increments arbitrarily determined is called the *independent variable*, and is always plotted on the x-axis. The other variable is termed the *dependent variable*, and is plotted on the y-axis. This dependence may be real, in the sense that

[1] It should be noted that letters at the end of the alphabet are used as symbols for *variables*, while letters at the beginning of the alphabet are used as symbols for *constants*, i.e. quantities the values of which do not change in the given discussion.

the values of the second variable are definitely determined by the values of the independent variable, or it may be purely the conventional dependence of the type described.

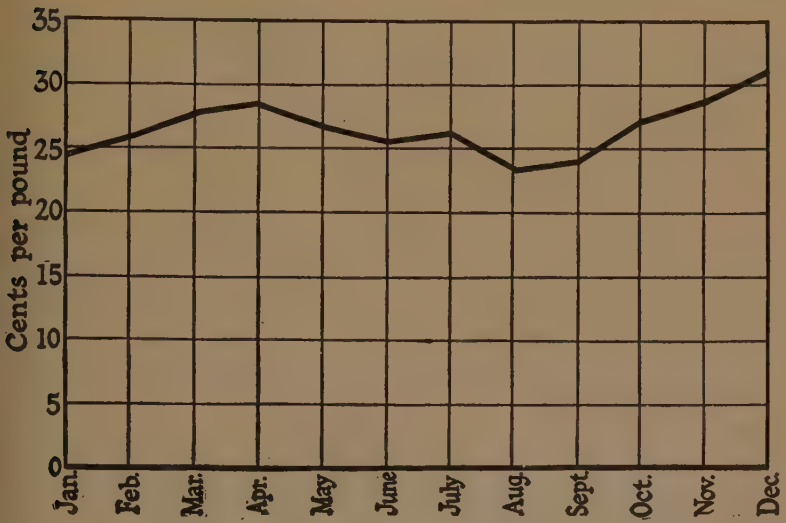


FIG. 2. — Cotton Prices in the United States, by Months, During the Year 1923
(Monthly averages of prices received by producers)

Time, it should be noted, is always plotted as independent, when it constitutes one of the variables.

FUNCTIONAL RELATIONSHIP

When the relationship between two variables is one of complete dependence, so that the value of y is uniquely determined by a given value of x, y is said to be a *function* of x. The general expression for such a relationship is $y = f(x)$. Thus the speed at which a body is falling at a given moment is a function of the time it has been falling, the pressure of a given volume of gas is a function of its temperature, the increase of a given principal sum of money at a fixed rate of interest is a function of time. If the values of the independent variable be laid off on the

x-axis of a rectilinear chart and the corresponding values of the function (i.e. the dependent variable) be laid off on the y-axis, a graphic representation of the function will be secured, in the form of a curve.[1] This concept of functional relationship is a very important one in statistical work. Some of the simpler functions may be briefly discussed.

The Straight Line

If two variables are so related that their values are always the same, their relationship is obviously of the form $y = x$. As a very simple example, the relation between the age of a tree and the number of rings in its trunk may be considered. A tree 20 years old will have 20 rings, and so on. This relationship may be represented on a coordinate chart, several sample values of x and y being taken. When these points are plotted and a line drawn through them, we secure a straight line passing through the origin and (assuming the two scales to be equal) bisecting the right angle XOY (cf. Fig. 3).

Similarly, any equation of the first degree, (i.e. not involving xy, or powers of x or y above the first) may be represented by a straight line. The generalized equation can be reduced to the form $y = a + bx$, where a is a constant representing the distance from the origin to the point of intersection of the given line and the y-axis, and b is a constant representing the slope of the given line (that is, the tangent of the angle which the line makes with the horizontal). The constant term a is called the *y-intercept*. It is clear from the generalized equation of the straight line that when x has a value of zero, y will be equal to this constant term. In the example given above (Fig. 3) a is equal to 0, and b to 1. The location of a given line depends upon the signs of a and b as well as upon

[1] The general term "curve" is used to designate any line, straight or curved, when located with reference to a coördinate system.

their magnitudes. The practical problem involved in the determination of any straight line is that of finding the values of a and b from the data, a problem which will

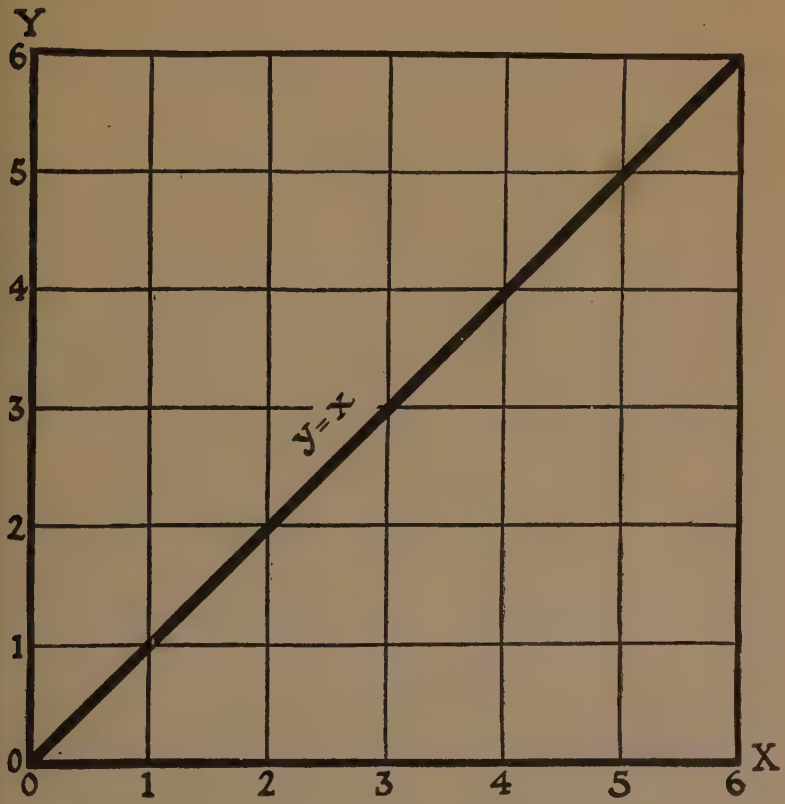


FIG. 3. — Graph of the Equation $y = x$

appear in various forms in the discussion of statistical methods.

These points may be illustrated by the plotting of a simple equation of the first degree. Thus, to construct the graph of the function, $y = 2 + 3x$, various values of x are assumed, and corresponding values of y are determined. These may be arranged in the form of a table:

x	y $(2 + 3x)$
− 4	− 10
− 2	− 4
0	2
2	8
4	14

Plotting these values and connecting the plotted points, the graph illustrated in Fig. 4 is secured. It will be noted that since this function is linear (that is, the graph takes the form of a straight line) any two of the points would have been sufficient to locate the line. The y-intercept is equal to the constant term 2, and the tangent of the angle which the given line makes with the horizontal (the slope of the line) is equal to 3, the coefficient of x. That this curve represents the equation is proved by the fact that the equation is satisfied by the coördinates of every point on the curve, and that every pair of values satisfying the equation is represented by a point on the curve. It is characteristic of a linear relationship that if one variable be increased by a constant amount, the corresponding increment of the other variable will be constant. In the above case as x grows by constant increments of two, for example, the constant increment of the y-variable is six. Series which increase in this way by constant increments are termed *arithmetic series*.

Many examples of linear relationship between variables are found in the physical sciences. An example from the economic world is found in the growth of money at simple interest, that is, interest which is not compounded. If we let r represent the rate of simple interest, x the number of years, and y the sum to which one dollar will amount at the end of x years, the equation of relationship is of the form

$$y = 1 + rx$$

Since in a given case r will be constant, this is of the simple linear type. In statistical work precise relationships of

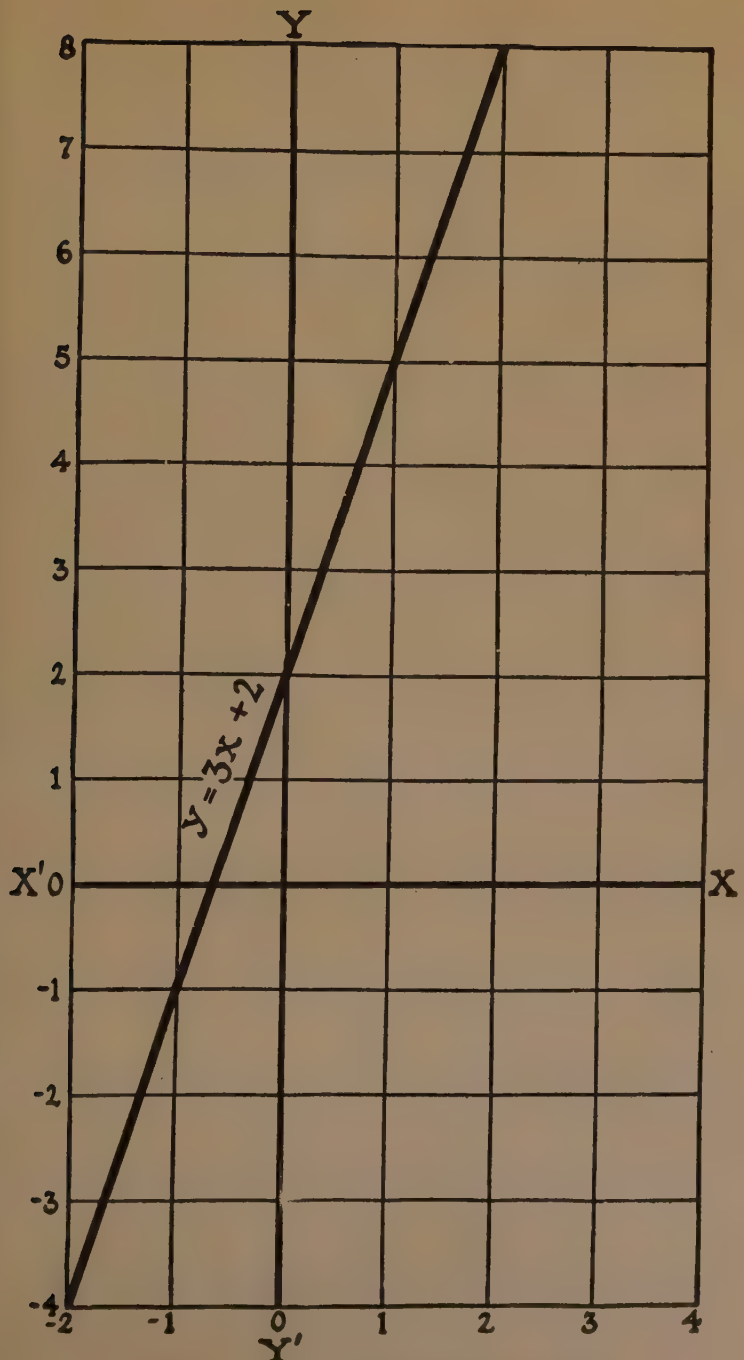


FIG. 4. — Graph of the Equation $y = 3x + 2$

this type rarely if ever occur, but approximations to the straight line relationship are found constantly.

Non-Linear Relationship

Non-linear functions are of many types, of which only a few of the more common will be discussed here. The student should be familiar with the general characteristics of the chief non-periodic curves, of which the parabolic and hyperbolic types, on the one hand, and the exponential type on the other, are the most important. The potential series is mentioned as a more general form of rather wide utility. Of periodic functions the sine curve is briefly described, as a fundamental form.

Functional relationships of the parabolic or hyperbolic form are quite common in the physical sciences, and such curves are found to fit certain classes of economic data. The general equation, when there is no constant term, is of the form $y = ax^b$. The curve is *parabolic* when the exponent b is positive, and *hyperbolic* when b is negative. The two following examples will serve to illustrate these types:

Problem: To construct the graph of the function $y = x^2$.

x	y (x^2)
− 5	25
− 4	16
− 3	9
− 2	4
− 1	1
0	0
1	1
2	4
3	9
4	16
5	25

The graph is shown in Fig. 5.

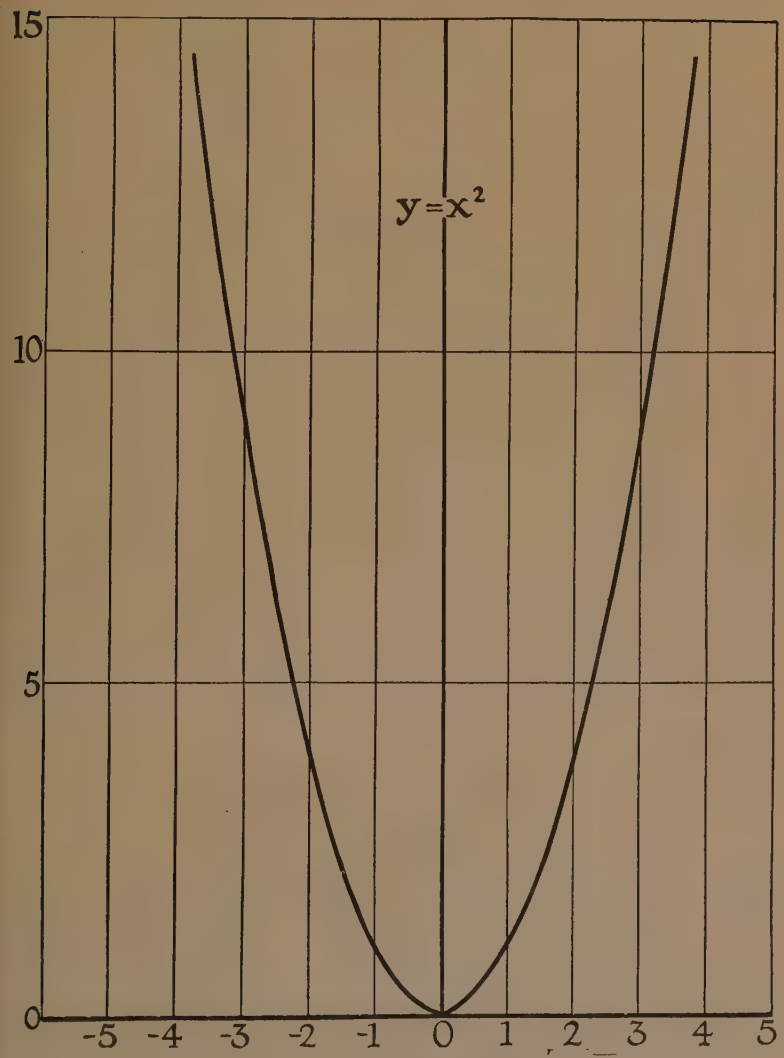


FIG. 5. — Parabola: Graph of the Equation $y = x^2$

Problem: To construct the graph of the function $y = x^{-1}$, for positive values of x.

x	y (x^{-1})
$\frac{1}{3}$	3
$\frac{1}{2}$	2
1	1
2	$\frac{1}{2}$
3	$\frac{1}{3}$
4	$\frac{1}{4}$
5	$\frac{1}{5}$

The graph of this function, an equilateral hyperbola, is shown in Fig. 6. It should be noted that this equation

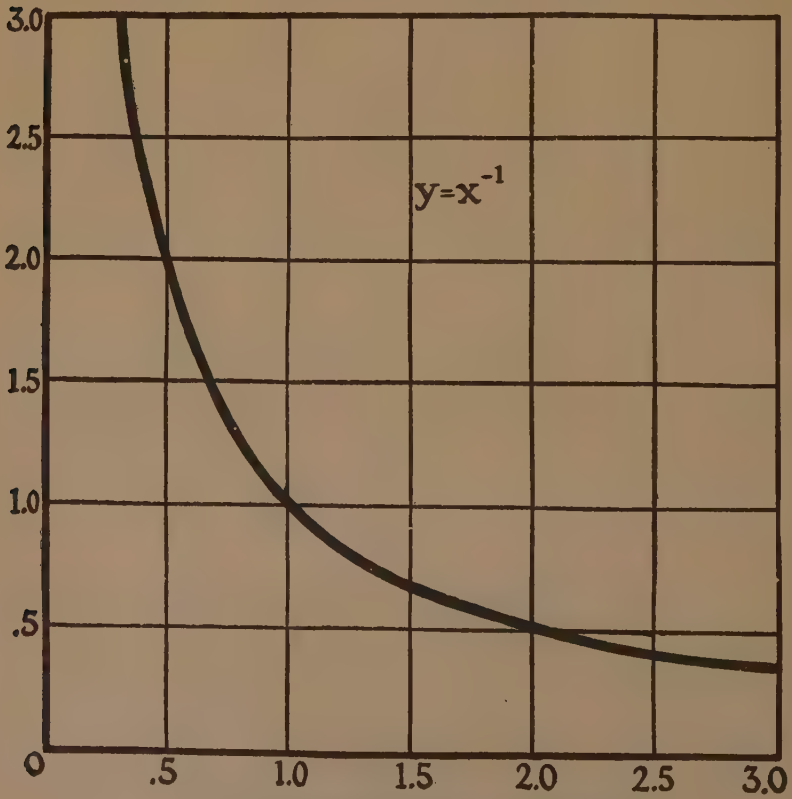


FIG. 6. — Equilateral Hyperbola: Graph of the Equation $y = x^{-1}$ (for positive values of x)

may also be written $y = \frac{1}{x}$ or $xy = 1$.

It is characteristic of relationships of this type that as x increases in geometric progression, y also increases in geometric progression. Thus, in the example of the parabola given above ($y = x^2$), if we select the x values which form a geometric series,[1] the corresponding y values form a similar series:

x	1	2	4	8	16	32
y	1	4	16	64	256	1024

Another class of functions is of the form represented by the equation $y = ab^x$. In equations of this type one of the variable quantities occurs as an exponent; graphs representing such equations are called *exponential curves.* The example which follows illustrates the type.

Problem: To construct the graph of the function $y = 2^x$, for positive values of x.

x	y (2^x)
0	1
1	2
2	4
3	8
4	16
5	32
6	64

This graph is shown in Fig. 7.

It has been noted that the relationship between two variables which increase by constant increments (constituting arithmetic series) may be represented by a straight line, and that the relationship between variables increasing in geometric progression may be represented by either a parabola or a hyperbola. The exponential curve constitutes a hybrid type. It describes a relation in which one variable increases in arithmetic progression while the other in-

[1] A geometric series is one each term of which is derived from the preceding term by the application of a constant multiplier.

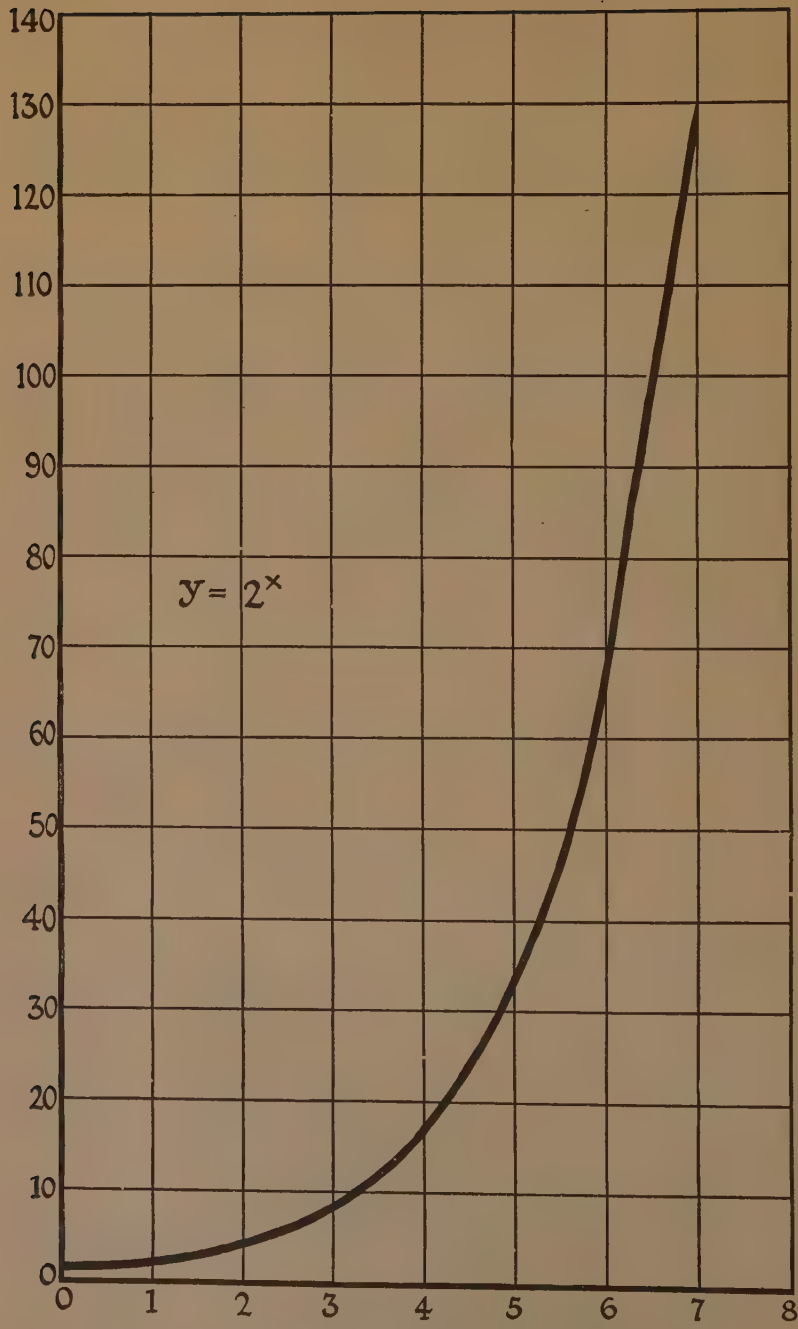


FIG. 7. — Exponential Curve: Graph of the Equation $y = 2^x$ (for positive values of x)

creases in geometric progression. The figures given above illustrate this relationship.

Curves based upon relationships of the following type have been employed extensively in statistical inquiries:

$$y = a + bx + cx^2 + dx^3 + \ldots\ldots$$

The term *potential series* has been applied to equations of this type. Though such curves do not constitute parabolas of the strict conic section type, a curve based upon such an equation carried to the second power of x is termed a second degree parabola, to the third power of x, a third degree parabola, etc. No uniform and simple type is secured from this series. It is treated in more detail at a later point.

Periodic functions constitute another distinct type, a class represented notably by electrical and meteorological relations, though not confined to these fields. The characteristic feature of such relationships is that values of the dependent variable repeat themselves at constant intervals of the independent variable. The sine curve, the basic type of this class, is illustrated in the following example.

Problem: To construct the graph of the function $y = \sin x$.

x (angle in degrees)	y ($\sin x$)
0°	.000
30°	.500
60°	.866
90°	1.000
120°	.866
150°	.500
180°	.000
210°	− .500
240°	− .866
270°	− 1.000
300°	− .866
330°	− .500
360°	.000
390°	.500
etc.	

The graph is shown in Fig. 8.

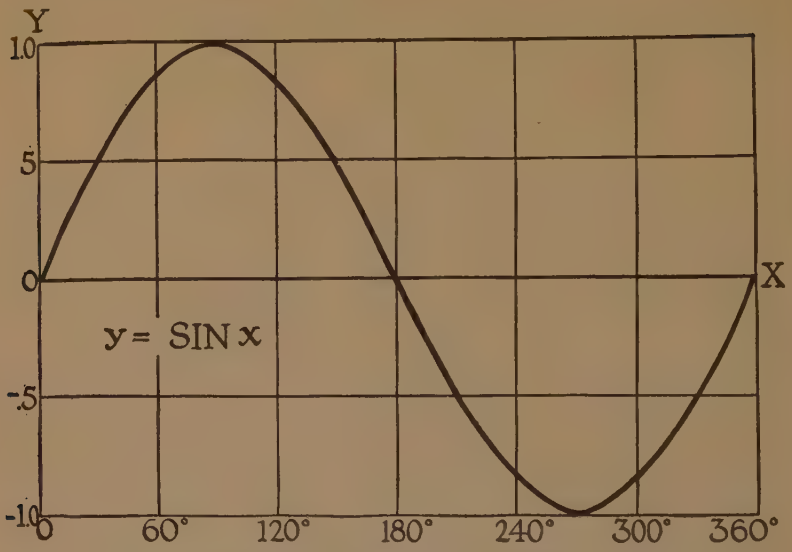


FIG. 8. — Sine Curve: Graph of the Equation $y = \text{Sin } x$

The full importance in statistical work of securing a mathematical expression for the relation between two variables cannot be demonstrated until the subject has been further developed. One fundamental object is the determination of physical or economic laws underlying observed phenomena. Another more practical object is the securing of a formula by means of which values of one variable may be approximated from given values of the other. Examples throughout the book will serve to illustrate how these objects are attained.[1]

LOGARITHMS

Logarithms, which play such an important part in general mathematical operations, are of equal importance in the manipulation of the raw materials of statistics.

[1] A fuller discussion of different curve types is presented below, in the section dealing with the analysis of time series.

The nature of logarithms, and the methods by which they are employed to facilitate arithmetic processes, may be briefly reviewed. This discussion is concerned only with the common system of logarithms of which the base is 10.

Any positive number may be expressed as a power of 10. Thus

$$1{,}000 = 10 \times 10 \times 10 = 10^3$$
$$10{,}000 = 10 \times 10 \times 10 \times 10 = 10^4$$

In each case the *exponent* of 10 (the small number written above and to the right) indicates the number of times the figure 10 is repeated as a factor. For the integral powers of 10 the exponent is a whole number, but for the other numbers the exponent will contain a fractional value. Thus 100 is equal to 10 raised to the power 2, or 10^2; 110 is equal to 10 raised to the power 2.04139, or $10^{2.04139}$.

The exponent of 10, or the index of the power to which 10 must be raised to equal a certain number, is called the *logarithm* of that number. The logarithm of 100 is 2, the logarithm of 110 is 2.04139, the logarithm of 998 is 2.99913. These figures all have reference to the base 10, though a system of logarithms might be developed on any base. In general, if

$$a = b^c$$
$$\log_b a = c$$

which may be read "the logarithm of a to the base b is equal to c." The relation between the given number, the base and the logarithm, when the common system of logarithms is employed, may be easily remembered if the following relations are kept in mind:

$$100 = 10^2$$
$$\log_{10} 100 = 2$$

The logarithm of any number has two parts, the integral and the decimal. The whole number is called the *characteristic*, and the decimal portion is termed the *mantissa*. The former is determined in a given case by inspection,

while the mantissa may be obtained from logarithmic tables. The characteristic varies with the location of the decimal point, while the mantissa remains the same for any given combination of numbers. This fact is illustrated by the following figures:

log of 8450	= 3.92686
log of 845	= 2.92686
log of 84.5	= 1.92686
log of 8.45	= .92686
log of .845	= 9.92686 − 10
log of .0845	= 8.92686 − 10

In finding the natural number to which a given logarithm corresponds (such natural numbers are termed *anti-logarithms*), the mantissa determines the sequence of figures, while the whole number, or characteristic, determines the location of the decimal point. For example, in seeking the anti-logarithm of 2.17609 it is found that the decimal .17609 follows the natural number 1500, in a table of logarithms. Since the characteristic is 2, the natural number desired must lie between 100 and 1000, and must therefore be 150.

A brief study of the following figures, showing the progression of numbers corresponding to certain powers of 10, will help to fix in mind the relations between the multiples of 10 and their logarithms, and will enable the characteristic of a desired logarithm to be readily determined.

.0001	.001	.01	.1	1	10	100	1,000	10,000
10^{-4}	10^{-3}	10^{-2}	10^{-1}	10^{0}	10^{1}	10^{2}	10^{3}	10^{4}

The exponents of 10 in the lower row are the logarithms of the numbers in the upper row.

It should be noted that the logarithms of all numbers from 0 to 1 are negative. Thus the logarithm of .845 is − 1 + .92686; this is written 9.92686 − 10. In covering the range of all positive natural numbers from zero to infinity, logarithms traverse all positive and negative

values. A negative natural number, therefore, can have neither a positive nor a negative logarithm.

The advantage of thus expressing numbers as powers of 10 lies in the fact that the ordinary arithmetic operations of multiplication, division, raising to powers and extracting roots are greatly facilitated by this procedure.

To multiply numbers, add their logarithms. The sum of the logarithms of the factors is the logarithm of their product. In general terms:

$$a^b \times a^c = a^{(b+c)}$$

Specifically

$$10^2 \times 10^3 = (10 \times 10) \times (10 \times 10 \times 10) = 10^5 = 100{,}000$$
$$100 \times 1{,}000 = 100{,}000$$

To divide one number by another, subtract the logarithm of the latter from the logarithm of the former. The remainder is the logarithm of the desired quotient.

In general terms:

$$a^b \div a^c = a^{(b-c)}$$

Specifically

$$10^5 \div 10^2 = \frac{10 \times 10 \times 10 \times 10 \times 10}{10 \times 10} = 10^3 = 1{,}000$$
$$100{,}000 \div 100 = 1{,}000$$

To raise a given number to any power, multiply the logarithm of the number by the index of the power. The product is the logarithm of the desired power.

In general terms:

$$(a^b)^c = a^{bc}$$

Specifically

$$(10^3)^2 = (10 \times 10 \times 10) \times (10 \times 10 \times 10) = 10^6 = 1{,}000{,}000$$
$$1{,}000^2 = 1{,}000{,}000$$

To extract any root of a given number, divide the logarithm of the number by the index of the root. The quotient is the logarithm of the desired root.

In general terms:

$$\sqrt[b]{a^c} = a^{\left(\frac{c}{b}\right)}$$

Specifically

$$\sqrt[3]{10^6} = 10^{\frac{6}{3}} = 10^2 = 100$$

$$\sqrt[3]{1{,}000{,}000} = 100$$

In summary:

$$\log (a \times b) = \log a + \log b$$
$$\log (a \div b) = \log a - \log b$$
$$\log a^b = b \times \log a$$
$$\log \sqrt[b]{a} = \log a \div b$$

These characteristic advantages of logarithms have been made use of in the construction of the slide rule, an instrument for reducing routine toil which should be familiar to all students of statistics.

Logarithmic Equations

The graphic representation of data by means of a system of rectangular coördinates has been described above and some of the advantages of this method have been outlined. For many purposes it is desirable to plot logarithms rather than the natural numbers themselves. This may result in bringing out significant relations more distinctly, or it may serve greatly to simplify and facilitate the manipulation of data. In particular, when it is possible through the use of logarithms to reduce a complex curve to the straight line form, a distinct gain has been made in the direction of simplicity of treatment and interpretation.

A linear equation, it will be recalled, is of the general form $y = a + bx$, where a and b are constants which measure, respectively, the y-intercept of the given line and the slope. The simplification of equations through the use of logarithms involves in all cases the substitution of $\log x$ or $\log y$, or both, for the x or y variables, thereby reducing an equation of a higher order to a simpler form.

This process may be illustrated with reference to the equation $y = x^2$. When plotted on rectangular coördinates this equation gives a curve of the parabolic type (cf. Fig. 5).

Reduced to logarithmic form this becomes log $y = 2 \log x$. This equation, in which the variables are log y and log x, is linear in form. It is plotted in Fig. 9, for positive values

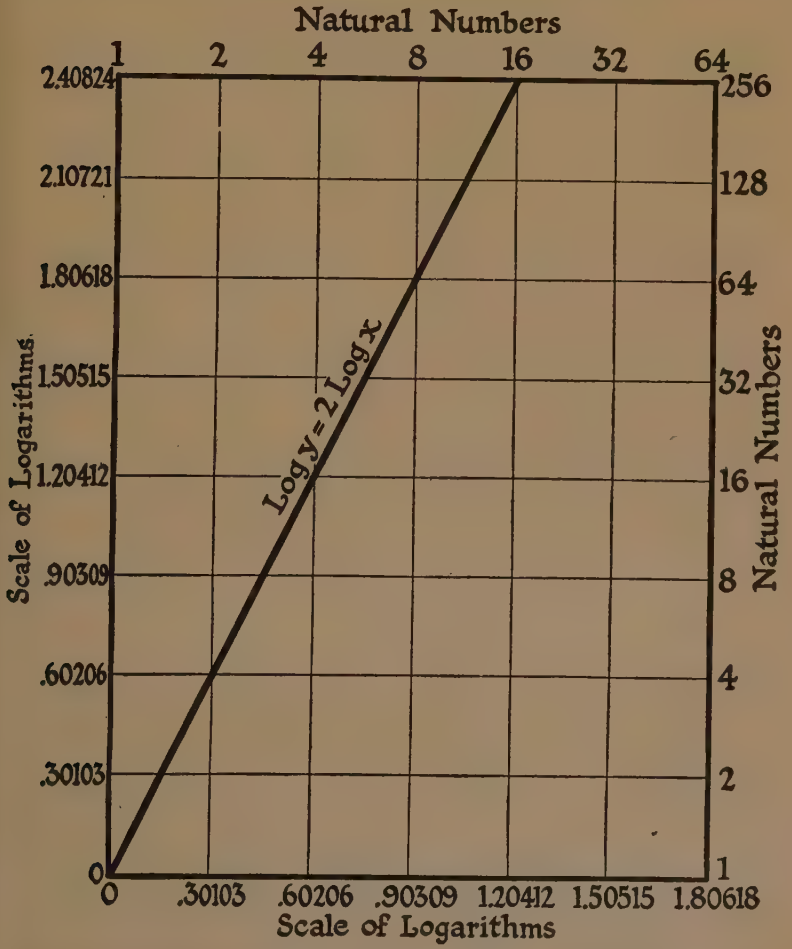


FIG. 9. — Graph of the Equation log $y = 2 \log x$. (Logarithmic form of the equation $y = x^2$)

of log x. To indicate the relations involved, natural numbers corresponding to the logarithms are given on scales to the right and at the top of the figure. The natural numbers

appearing on the scales constitute geometric series, while their logarithms form arithmetic series. Equal vertical distances on the chart, it will be noted, represent equal absolute increments on the scale of logarithms and equal percentage increments on the scale of natural numbers.

The equation $y = 5x^3$ can be reduced in the same way to $\log y = \log 5 + 3 \log x$, a linear form. Similarly, all equations of the type $y = ax^b$, that is to say, all simple parabolas and hyperbolas, can be reduced to the straight line form $\log y = \log a + b \log x$. Graphically this means plotting the logarithms of the y's against the logarithms of the x's.

A different problem is presented by an equation of the type $y = ab^x$, the graph of which is termed an exponential curve. Expressed in logarithmic form, we have $\log y = \log a + x \log b$. This is also of the linear type, the two constants being $\log a$ and $\log b$, while the variables are x and $\log y$. If we plot the natural x's and the logs of the y's, with this type of equation, a straight line will be secured. A curve of this type is discussed and illustrated below.

Logarithmic and Semi-Logarithmic Charts

There are certain disadvantages to the plotting of logarithms, however. If a considerable number of points are being plotted the task of looking up the logarithms may be tedious, and, in addition, the original values, in which chief interest lies, will not appear on the chart. These difficulties may be avoided by constructing charts with the scales laid off logarithmically, but with the natural numbers instead of the logarithms appearing on the scales. This is an arrangement identical with that employed in the construction of slide rules. Thus, although the natural numbers are given on the scales, distances are proportional to the logarithms of the numbers thereon plotted. In Fig. 10 such a chart is presented, showing the graph of the equation $y = x^2$.

A variation of this type of chart which is of great importance in statistical work is one which is scaled arithmetically on the horizontal axis and logarithmically on the vertical axis. This is equivalent, of course, to plotting the

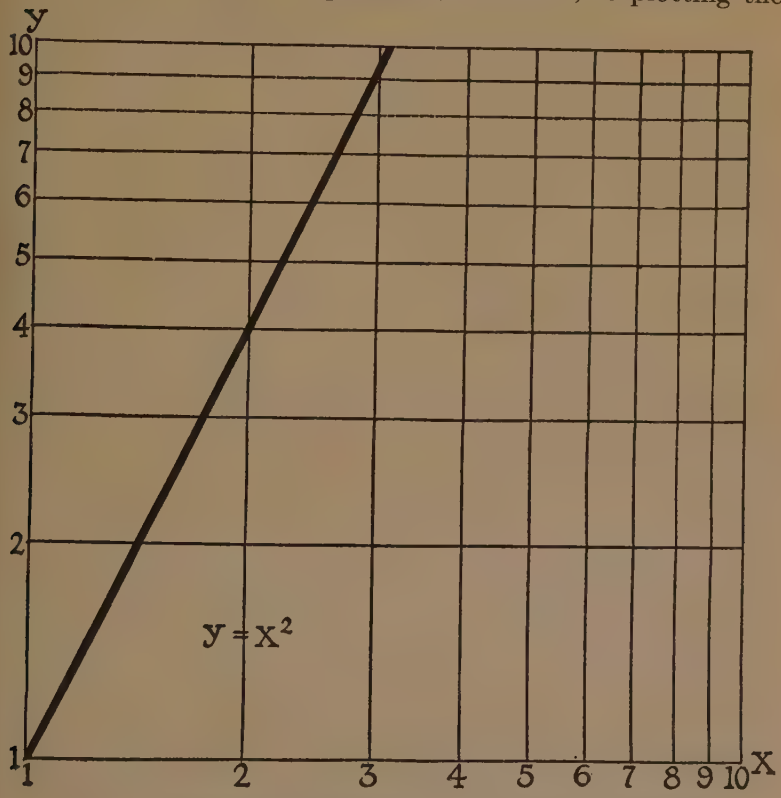


FIG. 10. — Graph of the Equation $y = x^2$. (Plotted on paper with logarithmic scales)

x's on the natural scale and plotting the logarithms of the y's. As was pointed out above, such a combination of scales reduces a curve of the exponential type to a straight line. Plotting paper of this semi-logarithmic or "ratio" type may be constructed with the aid of a slide rule or of logarithms, or may be purchased ready made. It is of particular value in charting economic statistics, because of

the fact that time is usually one of the variables in such cases, and it is desirable to plot this variable on the natural scale.

As an example of this type of curve the compound inter-

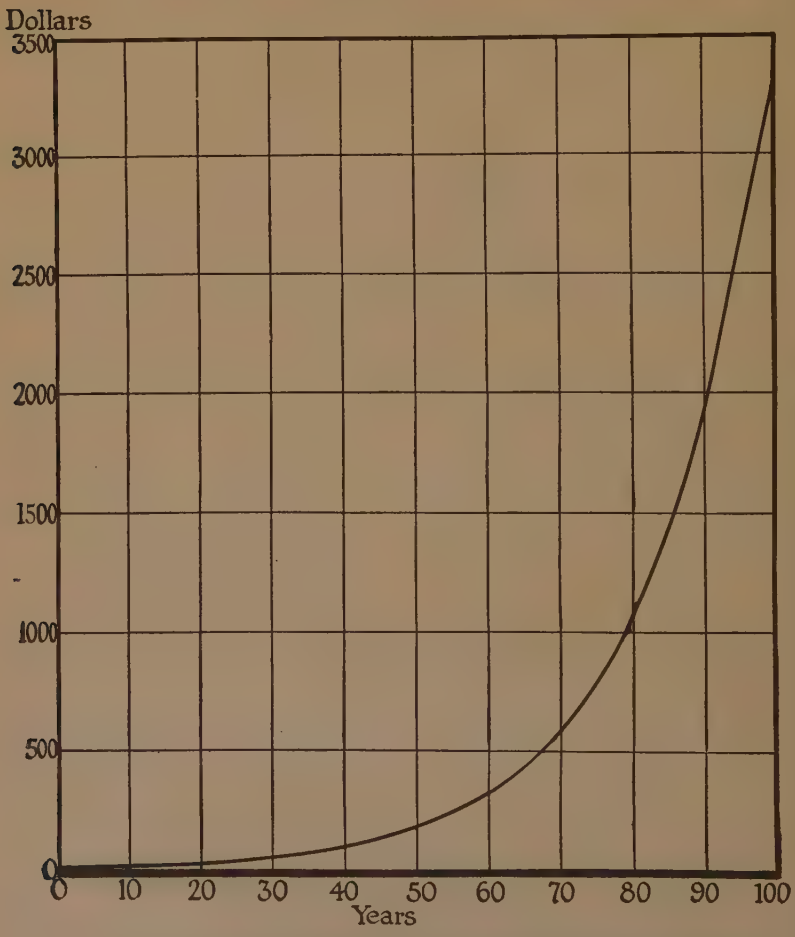


FIG. 11. — The Compound Interest Law: The Growth of $10.00 at Compound Interest at 6% for 100 Years. (Plotted on the arithmetic scale)

est law may be used. If r be taken to represent the rate of interest, x the number of years, p the principal, and y the sum to which the principal amounts at the end of x

years (interest being compounded annually), an equation is secured of the form

$$y = p(1 + r)^x$$

Expressed logarithmically this becomes

$$\log y = \log p + x \log (1 + r)$$

the equation to a straight line.

In Fig. 11 a curve representing the growth of ten dollars

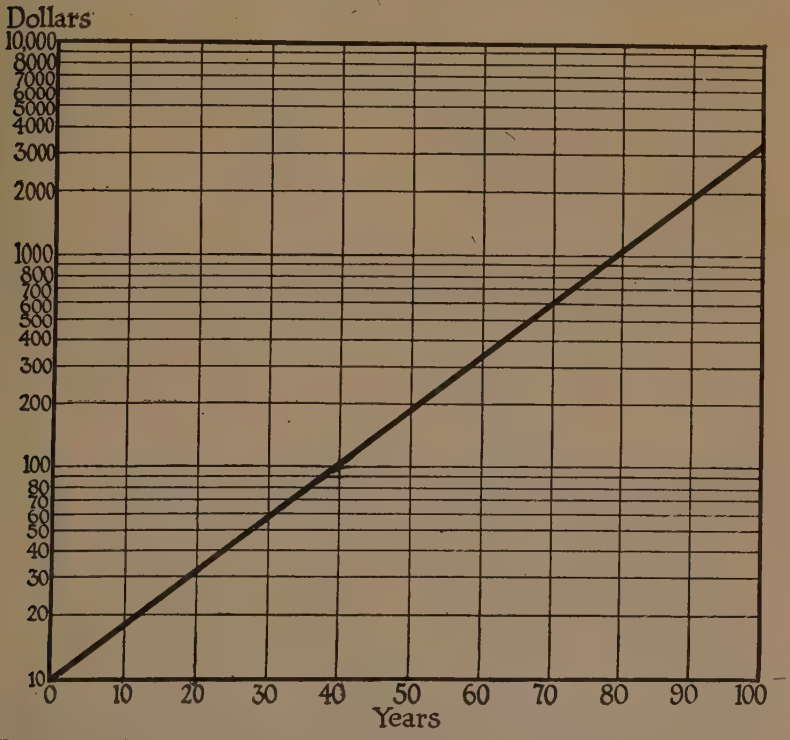


FIG. 12. — The Compound Interest Law: The Growth of $10.00 at Compound Interest at 6% for 100 Years. (Plotted on the semi-logarithmic or ratio scale)

at compound interest at 6 per cent is plotted on the natural scale. This is the graph of the exponential equation

$$y = 10\ (1 + .06)^x$$

y representing the total amount of principal and interest at the end of x years. Figure 12 shows the same data

plotted on semi-logarithmic paper, the exponential curve being reduced to a straight line.

The use of semi-logarithmic paper is not confined to cases in which an exponential curve is straightened out, for the significance of many types of data is most effectively brought out when charts of this type are used. These advantages are more fully explained below.

The Construction of Charts

When the results of observations or statistical investigations have been secured in quantitative form, one of the first steps toward analysis and interpretation of the data is that of presenting these results graphically. Not only is such procedure of scientific value in paving the way for further investigation of relationships, but it serves an immediate practical purpose in visualizing the results. A visual stimulus opens up a far more direct path to our understanding and imagination than that afforded by the more recently developed processes of reasoning. The interpretation of a column of raw figures may be a difficult task; the same data in graphic form may tell a simple and easily understood story. For these reasons graphic methods of presentation have come to play a highly important part in the everyday activities of business, as well as in the laboratory and drafting room.

It is beyond the scope of this book to present any detailed account of the multiplicity of graphs employed by engineers and statisticians today. Certain of the more important principles of graphic presentation may be briefly explained, however, and some of the chief types of graphs which are in daily use may be illustrated. Other examples appear in later chapters of this book.

Factors Governing the Selection of a Chart

The selection of the type of chart to be employed in a given case will depend upon two general considerations.

The first of these relates to the character of the material to be plotted. While the data of a given problem may frequently be presented graphically in several different forms, there is generally one type of chart best adapted to that material. It may be true, also, that certain types would be quite inappropriate to the data in question. The selection of a type of chart to employ, therefore, must be made with the characteristics of the data clearly in mind.

Perhaps more important is the *purpose* which the given chart is designed to serve. Each of the many types of charts in common use is appropriate to certain specific purposes. It will bring out certain characteristics of the data or will emphasize certain relationships. There is no chart which is sovereign for all purposes. Until the purpose is clearly defined the best chart form can not be selected. The following descriptions of a few standard types will facilitate the selection of an appropriate form.

Charts Adapted to the Plotting of Time Series

In the graphic presentation of a time series, primary interest attaches to the chronological variations in the values of the data, to the general trend and to the fluctuations about the trend. If the purpose is to emphasize the *absolute* variations, the differences in absolute units between the values of the series at different times, a simple chart of the type illustrated in Fig. 13 will serve the purpose. This chart depicts annual wheat flour exports from the United States during the period 1901–1923. Both scales are arithmetic. Points representing the various annual values are shown and, to facilitate interpretation, these points are connected by a series of straight lines. The chart tells a simple story of year-to-year fluctuations, with a general downward trend to 1910, an upward trend from 1911 to 1919, and a decline since 1919. With respect to general make-up, the following points should be noted:

1. The title constitutes a clear description of the material plotted and indicates the period covered.
2. The vertical scale begins at the zero line, enabling a true impression to be gained of the magnitude of the fluctuations.
3. The zero line and the line joining the plotted points are ruled more heavily than the coördinate lines.

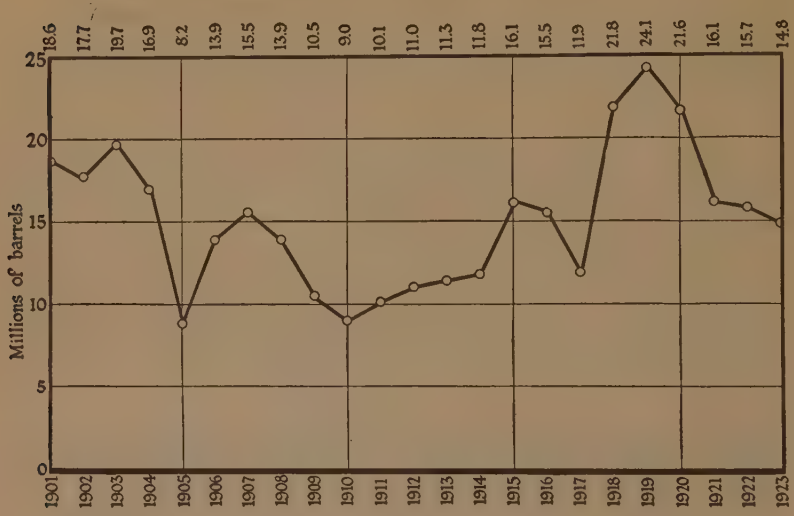


FIG. 13. — Wheat Flour Exports from the United States, 1901–1923

4. Figures for the scales are placed at the left and at the bottom of the chart. The vertical scale may be repeated at the right to facilitate reading. All figures are so placed that they may be read from the base as bottom or from the right hand edge of the chart as bottom.
5. The y-values of the plotted points are given at the top of the chart. This practice is helpful, though not necessary, as the values may be presented in a separate table.

ADVANTAGES OF THE RATIO CHART

If *relative* rather than *absolute* variations are of chief concern, the chart employed should be of the semi-logarithmic type, scaled logarithmically on the y-axis and arith-

metically on the x-axis. In such a chart equal percentage variations are represented by equal vertical distances, as opposed to the ordinary arithmetic type in which equal absolute variations are represented by equal vertical distances. The argument for the use of the semi-logarithmic or ratio chart for the representation of time series is that,

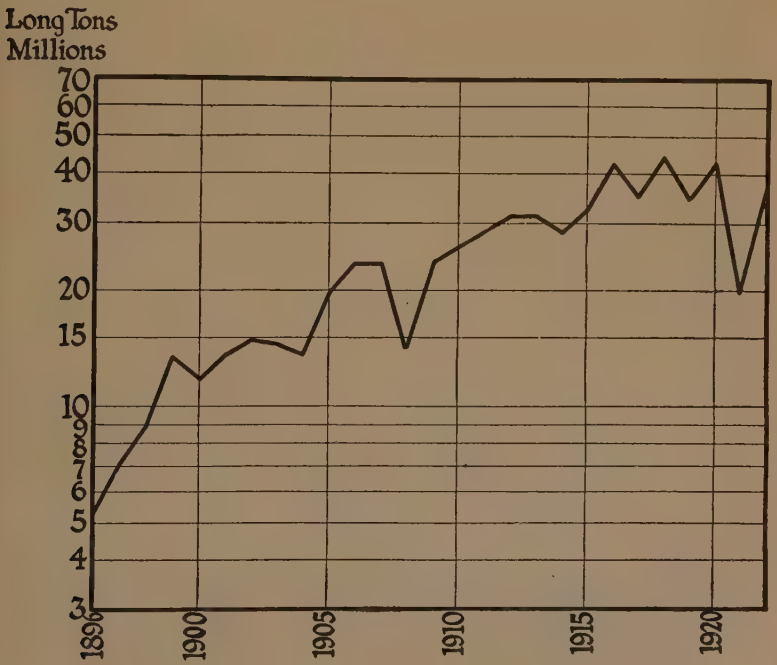


FIG. 14. — Production of Steel in the United States, 1896–1922
(Plotted on the semi-logarithmic scale)

in general, the significance of a given change depends upon the magnitude of the base from which the change is measured. That is, an increase of 100 on a base of 100 is as significant as an increase of 10,000 on a base of 10,000. In each case there is an increase of 100%. The absolute increase in the second case is 100 times that in the first case, and the two changes would show in this proportion on the

arithmetic chart. They would show as of equal importance on the semi-logarithmic chart.

Such a chart is presented in Fig. 14, which shows the course of steel production in the United States from 1896 to 1922. The absolute magnitudes are plotted, but the vertical scale is so constructed as to represent variations from year to year in proportion to their relative magnitude.

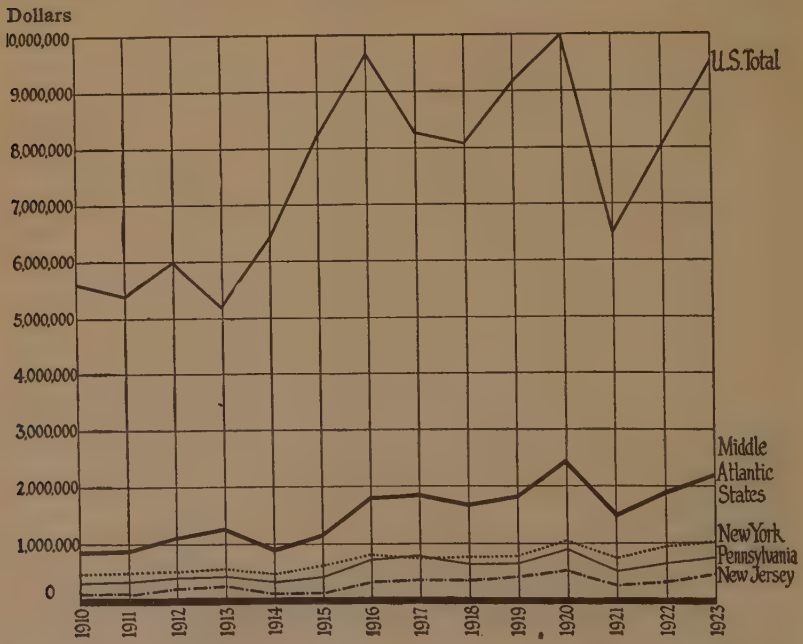


FIG. 15. — Sales of the Acme Corporation, 1910–1923. Showing the Total Sales in the United States and the Sales in Certain Subdivisions. (Plotted on the arithmetic scale)

Certain distinctive advantages of the ratio or logarithmic ruling are brought out by a comparison of Fig. 15 and Fig. 16. The data presented graphically in these two charts are the following:

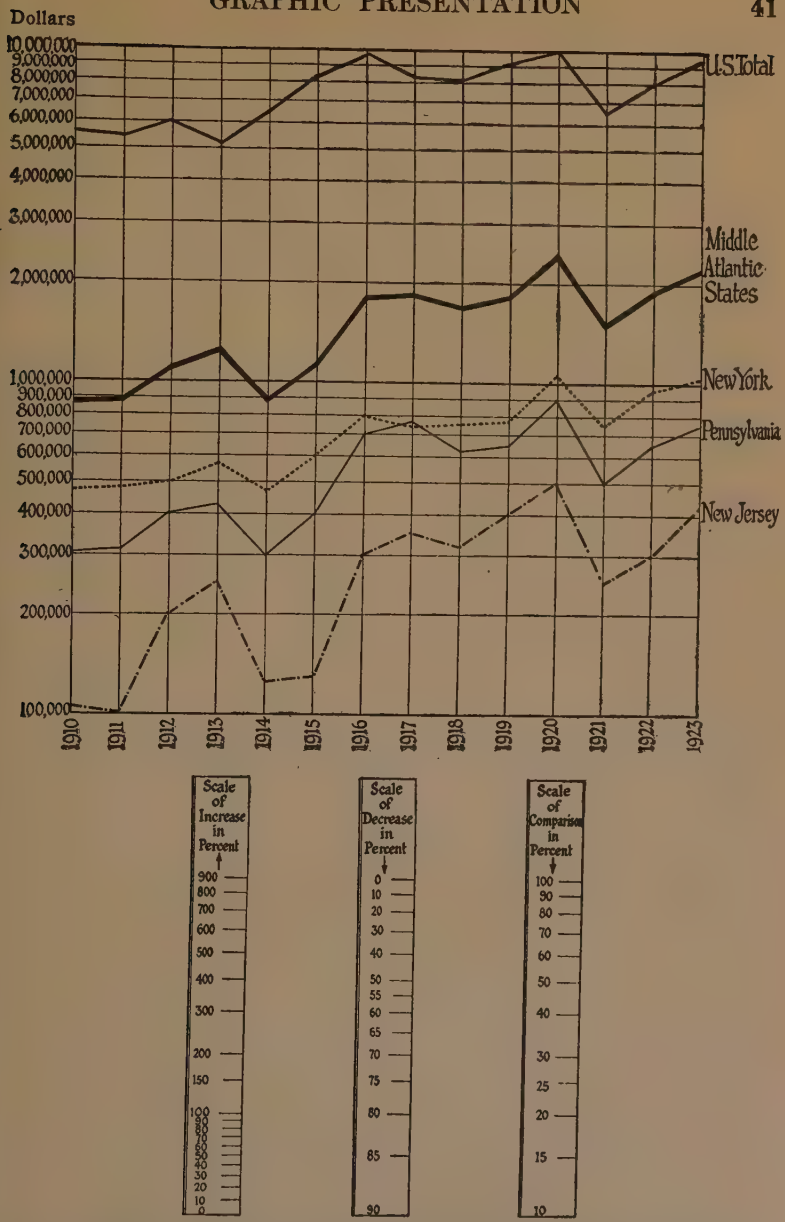


FIG. 16. — Sales of the Acme Corporation, 1910–1923. Showing the Total Sales in the United States and the Sales in Certain Subdivisions. With Scales of Increase, Decrease and Comparison. (Plotted on semi-logarithmic paper)

TABLE 2

Sales of the Acme Corporation, 1910–1923

	Penn.	*New Jersey*	*New York*	*Total Middle Atlantic States*	*Total United States*
1910.....	$305,000	$105,000	$465,000	$ 875,000	$5,600,000
1911.....	310,000	100,000	480,000	890,000	5,400,000
1912.....	400,000	200,000	500,000	1,100,000	6,000,000
1913.....	425,000	250,000	575,000	1,250,000	5,200,000
1914.....	300,000	125,000	465,000	890,000	6,400,000
1915.....	400,000	130,000	600,000	1,130,000	8,200,000
1916.....	700,000	300,000	800,000	1,800,000	9,700,000
1917.....	760,000	350,000	740,000	1,850,000	8,300,000
1918.....	630,000	320,000	750,000	1,700,000	8,100,000
1919.....	650,000	400,000	775,000	1,825,000	9,200,000
1920.....	900,000	500,000	1,050,000	2,450,000	10,000,000
1921.....	500,000	250,000	750,000	1,500,000	6,500,000
1922.....	650,000	300,000	950,000	1,900,000	8,000,000
1923.....	750,000	425,000	1,025,000	2,200,000	9,500,000

If the five series are to be presented on a single chart, scaled arithmetically, a scale must be selected which will include the largest item recorded, $10,000,000. Such a scale reduces the relative importance of the smaller magnitudes. From Fig. 15 it appears that during the period covered by the chart very considerable fluctuations occurred in the sales for the country as a whole, that minor fluctuations occurred in the total for the Middle Atlantic States, and that the sales in the three individual states represented remained practically constant. Such a picture is quite misleading. The true state of affairs is reflected in Fig. 16, in which the same data are plotted on paper with a semi-logarithmic ruling. The fluctuations in the individual states now appear to have been relatively greater than those in the total for the country as a whole. For the purpose of comparing series which differ materially with respect to the magnitude of the individual items, the arithmetic ruling is quite useless, giving a thoroughly distorted picture of

the true relations. The ratio ruling permits a legitimate comparison.

The scales printed below Fig. 16 emphasize certain very useful features of the logarithmic ruling. The *scale of increase* may be used to measure with a fair degree of accuracy the increase in a given series between any two dates. A given vertical distance on the chart, it will be recalled, represents a constant percentage increase at all points on the chart. Thus the distance from $100,000 to $200,000, along the vertical scale, is the same as the distance from $2,000,000 to $4,000,000. Any vertical distance may be measured, and the percentage of increase which it represents may be determined by laying off the given distance along the scale of increase, which is always read from the bottom up. For example, to determine the degree of increase in New Jersey sales from 1912 to 1913, we measure the vertical distance between the points plotted for these two years. Laying off this distance along the scale, it is found to represent a 25 per cent increase.

The *scale of decrease* is used in a similar fashion. The vertical distance between any two points is measured, and the percentage decrease which it represents is determined by laying off the given distance on the scale from the top downward. The arrows indicate the direction in which the various scales are to be read.

By means of the *scale of comparison* the percentage relation of one series to another at any time may be determined. For example, we may wish to know the percentage relation between Middle Atlantic sales and total sales in the United States in 1912. The vertical distance between the two plotted points is measured, and laid off on the scale of comparison, reading from the top downward. It is found to be approximately 18 per cent.

Scales of the type illustrated above may be readily constructed on a given chart by using the ratio ruling for the scale intervals. When a series of charts is constructed on

semi-logarithmic paper of a standard type it is more convenient to construct such scales in a more permanent form, in the shape of special rulers.

The chief advantages of the semi-logarithmic ruling in chart construction may be briefly summarized:

1. A curve of the exponential type becomes a straight line when plotted on a semi-logarithmic chart. For example, a curve representing the growth of any sum of money at compound interest takes the form of a straight line when so plotted.
2. In any series, so long as the *rate* of increase or decrease remains constant the graph will be a straight line on this ruling.
3. Equal relative changes are represented by lines having equal slopes. Thus two series increasing or decreasing at equal rates will be represented by parallel lines.
4. Comparison of the rates of change in two or more series is effected by comparison of the slopes of the plotted lines.
5. The semi-logarithmic ruling permits, at the same time, the plotting of absolute magnitudes and the comparison of relative changes.
6. Comparison of series differing materially in the magnitude of individual items is possible with the semi-logarithmic chart.
7. Percentages of change may be read and percentage relations between magnitudes determined directly from the chart.

Charts for the Comparison of Frequencies

A different type of chart is called for when the object is the comparison of frequencies, that is, numbers of events or things of different classes. The following census figures may serve to illustrate the problem.

Table 3

Farms in New England States in 1920

State	*Number of farms*
Maine	48,227
New Hampshire	20,523
Vermont	29,075
Massachusetts	32,001
Rhode Island	4,083
Connecticut	22,655

A graphic comparison of these six states with respect to number of farms in 1920 is afforded by the bar diagram in Fig. 17. This is a simple but effective type of chart for this purpose.

Further examples of this type of chart, as employed in the representation of frequency distributions, are contained

FIG. 17. — Farms in New England States in 1920

in the next chapter. It is there shown how a frequency polygon or frequency curve may grow out of the simple bar diagram, when data of certain kinds are being handled. Such frequency curves constitute very important graphic types, but it will be more appropriate to treat them in full at a later point.

CHARTS FOR THE REPRESENTATION OF COMPONENT PARTS

It is frequently desirable in graphic presentation to break up a total into its component parts, in order that changes in the parts as well as in the total may be followed. The following table constitutes a simple example of such data.

TABLE 4

Production Costs of the XYZ Corporation, 1923

(The figures show the total and subdivided cost per unit produced)

Month	*Material Cost*	*Labor Cost*	*Overhead Cost*	*Total Cost*
January	$32.00	$12.00	$6.00	$50.00
February	31.00	11.00	6.00	48.00
March	33.00	12.00	6.50	51.50
April	31.00	13.00	6.00	50.00
May	27.00	13.00	7.00	47.00
June	24.50	13.50	7.00	45.00
July	23.00	16.00	7.00	46.00
August	22.00	18.50	8.00	48.50
September	23.50	18.00	7.50	49.00
October	24.00	16.00	7.50	47.50
November	24.00	17.00	8.00	49.00
December	24.50	17.50	8.50	50.50

These figures are presented graphically in Fig. 18. It is clear from this diagram that, while total costs have been relatively stable, certain of the elements in this total cost have fluctuated quite widely. A knowledge of the total alone is inadequate, and must be supplemented by knowledge of the changes in the component items. Such a chart as this presents a clear picture, not only of changes in the total, but of the movements of each of the constituent parts.

CUMULATIVE CHARTS

In many cases chief interest in the development of a series attaches not to the value of each successive item but to the cumulated total of a number of such items. This may be so when a yearly production program has been laid out. In such a case it is the relation between cumulated production to date and scheduled production to date which is of major interest, and a chart form is needed which will enable this comparison to be made. The following figures illustrate the type of data for which such charts are appropriate.

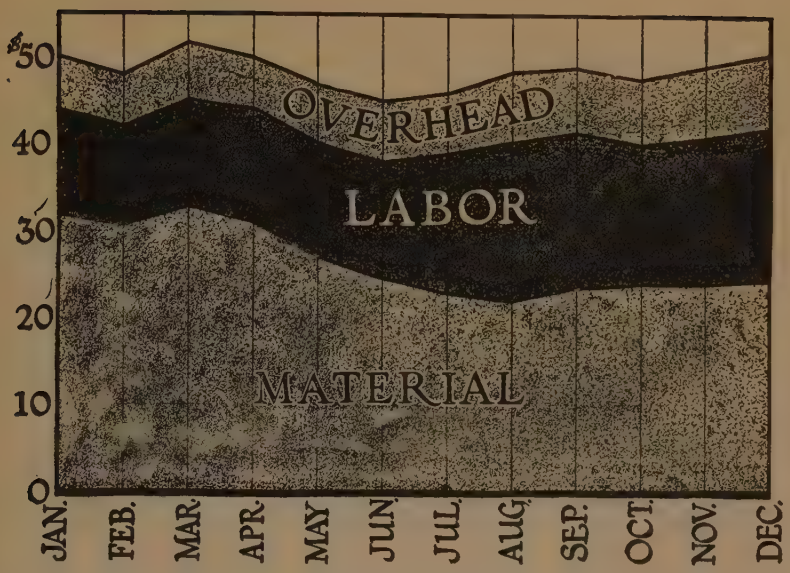


FIG. 18. — Elements of Production Costs of the XYZ Corporation By Months, 1923

TABLE 5

Cumulative Production Schedule and Cumulative Output, 1924

Speedwell Automobile Company

Month	*Production Schedule (cars)*	*Cumulative Production Schedule (cars)*	*Output (cars)*	*Cumulative Output (cars)*
January	2,000	2,000	750	750
February	2,000	4,000	1,250	2,000
March	2,000	6,000	1,250	3,250
April	3,000	9,000	2,000	5,250
May	3,000	12,000	1,500	6,750
June	3,750	15,750	3,750	10,500
July	3,750	19,500	4,250	14,750
August	5,000	24,500	6,000	20,750
September	4,000	28,500	5,750	26,500
October	4,000	32,500		
November	4,000	36,500		
December	4,000	40,500		

It is assumed that this table represents the situation as of the end of September.

In Fig. 19 the two cumulative curves are plotted. The relation between actual and scheduled production at the

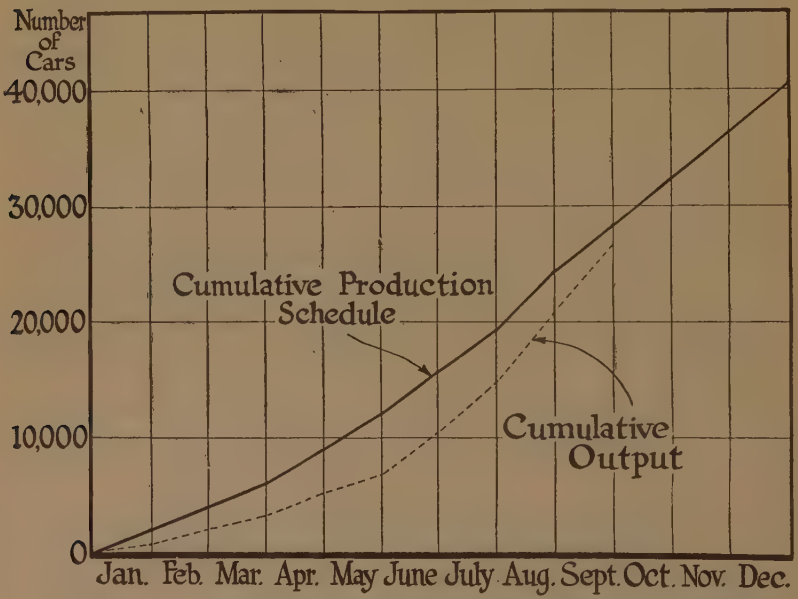


FIG. 19. — Comparison of Scheduled and Actual Output (Cumulative) Speedwell Automobile Company, 1924

end of each month is shown on the chart, and it is possible from the scale to read the approximate amount by which production is behind schedule. By reference to the figures, which should always accompany the chart, the exact relation may be determined. Such a chart has many applications, some of which are illustrated in the following chapter.

THE GANTT PROGRESS CHART

The same data may be presented in a very effective form by making use of a type of chart developed by Mr. H. L. Gantt. An adequate description of this chart and of its

many uses would far exceed the space which can be given to it here, but its characteristics may be indicated in a very brief account.

Once a schedule has been drawn up, the Gantt chart may be utilized in checking actual accomplishment against the schedule. Having such a schedule as that given in Table 5, the monthly and annual quotas may be entered on a form similar to that shown in Fig. 20. The entry to the left of each monthly space indicates the amount scheduled for production during that month. The entry to the right of each monthly space indicates the cumulated scheduled production to the end of the given month. In this figure the results of the first two months' operations are shown. The heavy black line indicates the cumulated actual production during this period, amounting to 2000 cars. The narrow upper lines in the January and February columns measure the actual production in each of those months. If actual production in either month had equaled the scheduled production the light line would extend across the full monthly space. When actual production in a given month exceeds the scheduled production a double light line appears.

It should be noted that the spaces into which each monthly period is divided represent equal time intervals but varying amounts in terms of actual production. Thus the space representing one fifth of the January interval represents a production of 400 cars (the January quota being 2000). The space representing one fifth of the August interval represents 1000 cars (the August quota being 5000). In reading the chart in terms of absolute magnitudes reference must be had to the monthly quotas.

The situation at the end of September is shown in Fig. 21. The arrow at the top of the diagram indicates the point of time actually reached. That actual production is behind scheduled production by one-half month is apparent from the relation between this arrow and the heavy

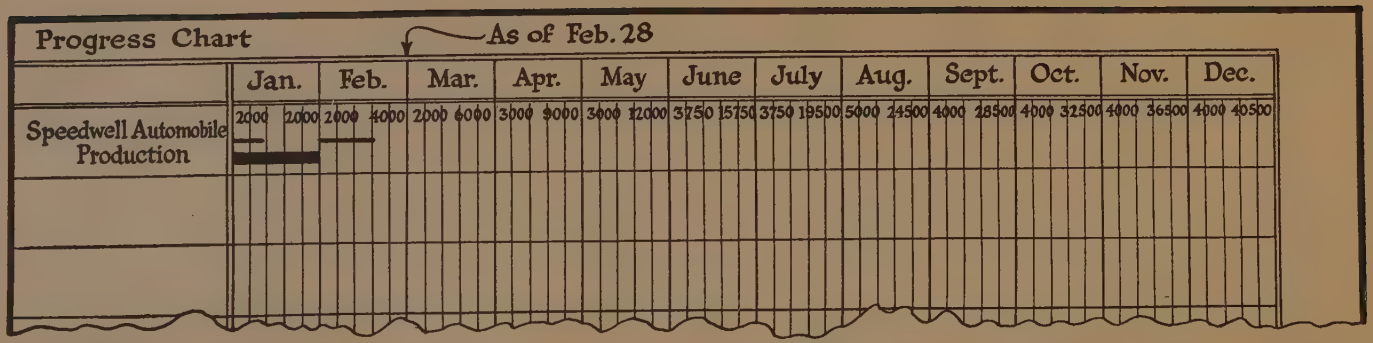


FIG. 20. — Comparison of Scheduled and Actual Output: Gantt Progress Chart. (Showing the situation on February 28th.)

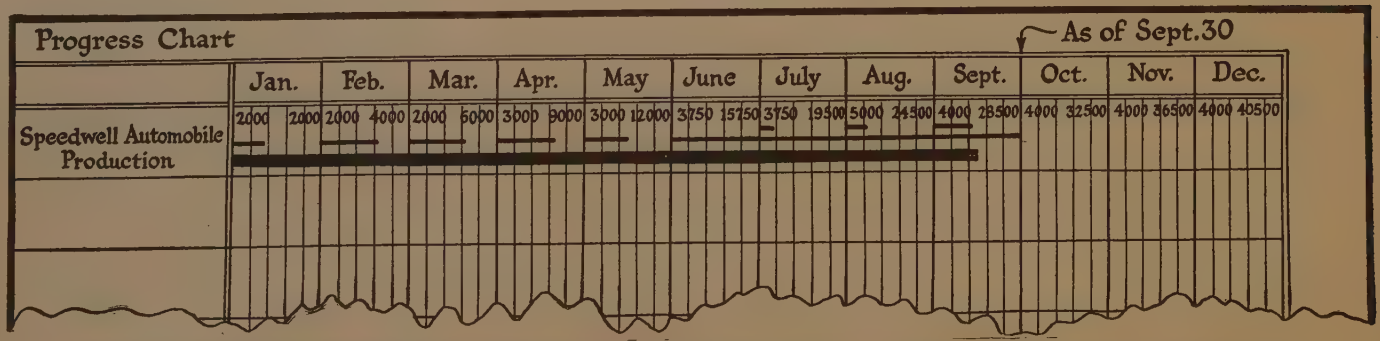


FIG. 21. — Comparison of Scheduled and Actual Output: Gantt Progress Chart. (Showing the situation on September 30th.)

black line, while the light lines indicating monthly production show that actual output has exceeded the monthly quota in each of the last three months.

The Gantt chart has a great variety of applications in governmental and business organizations. The economy of space is such that developments in a number of departments or districts may be shown on a single chart. It constitutes the simplest and most effective graphic method known for following the progress of work under way, for comparing actual accomplishment with an established program. And in so doing, it increases by so much the efficiency of administrative control.

Standard Rules for Graphic Presentation

Graphic methods have been widely employed in the physical and social sciences and in business, and the resulting diversity of uses has made it difficult to secure standardization of practice. To remedy this defect, a joint committee composed of representatives of the various groups interested in this subject prepared a report recommending the employment of certain standard methods in graphic work. This report, which is a model of clear and succinct presentation, has appeared in various publications, but it cannot be too widely circulated. The suggestions of the committee are presented below [1]:

[1] This report, which is the work of the Joint Committee on Standards for Graphic Presentation, of which Willard C. Brinton was chairman, appeared in the *Quarterly Publications of the American Statistical Association*, Vol. 14, 790–797, 1915. It is presented here with the permission of Mr. Brinton. Copies may be secured from the American Society of Mechanical Engineers, New York.

1. The general arrangement of a diagram should proceed from left to right.

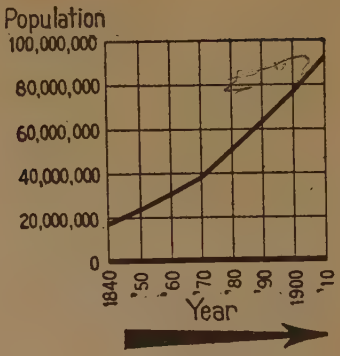


Illustration 1

Illustration 2

2. Where possible represent quantities by linear magnitudes, as areas or volumes are more likely to be misinterpreted.

3. For a curve the vertical scale, whenever practicable, should be so selected that the zero line will appear on the diagram.

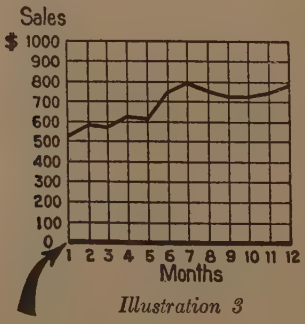


Illustration 3

4. If the zero line of the vertical scale will not normally appear on the curve diagram, the zero line should be shown by the use of a horizontal break in the diagram.

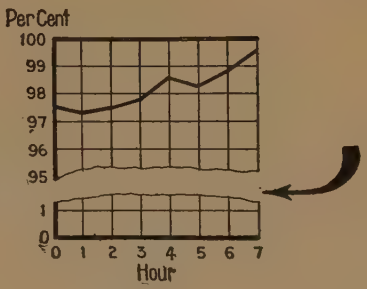


Illustration 4

5. The zero lines of the scales for a curve should be sharply distinguished from the other coördinate lines.

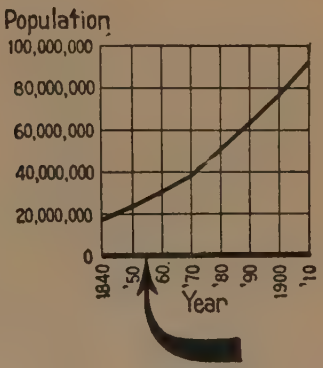


Illustration 5a

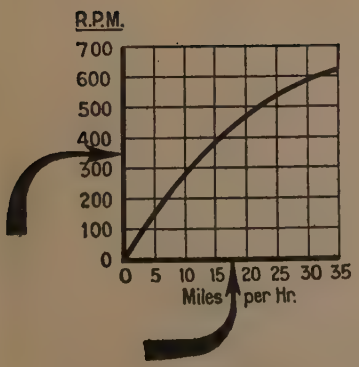


Illustration 5b

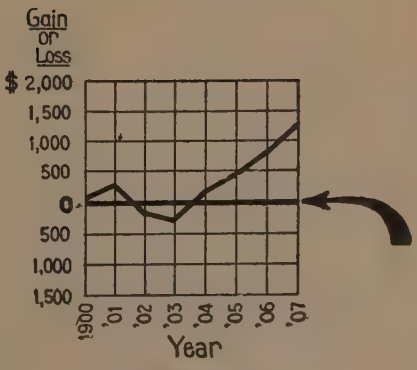


Illustration 5c

6. For curves having a scale representing percentages, it is usually desirable to emphasize in some distinctive way the 100 per cent line or other line used as a basis of comparison.

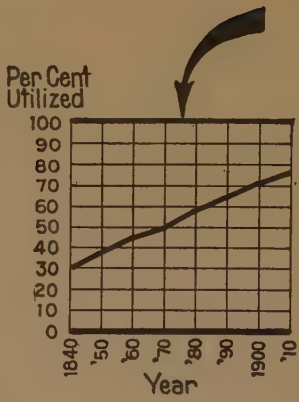


Illustration 6a

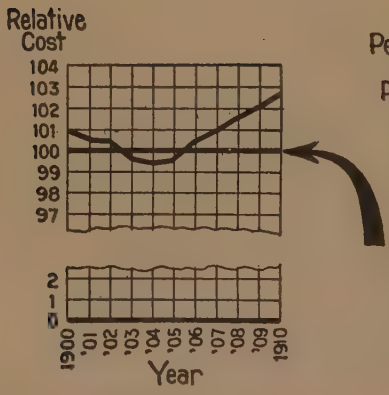


Illustration 6b

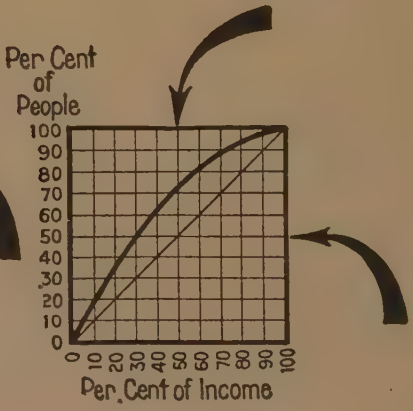


Illustration 6c

7. When the scale of a diagram refers to dates, and the period represented is not a complete unit, it is better not to emphasize the first and last ordinates, since such a diagram does not represent the beginning or end of time.

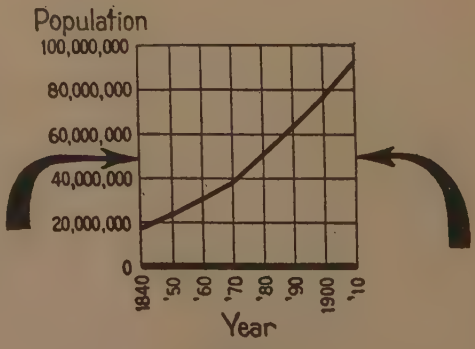


Illustration 7

8. When curves are drawn on logarithmic coördinates, the limiting lines of the diagram should each be at some power of ten on the logarithmic scales.

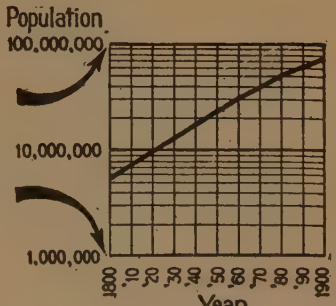


Illustration 8

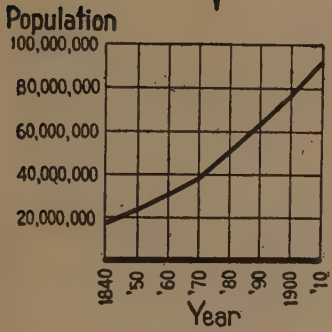


Illustration 9a

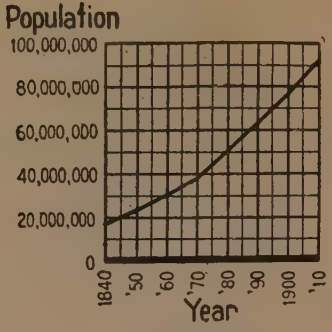


Illustration 9b

9. It is advisable not to show any more coördinate lines than necessary to guide the eye in reading the diagram.

10. The curve lines of a diagram should be sharply distinguished from the ruling.

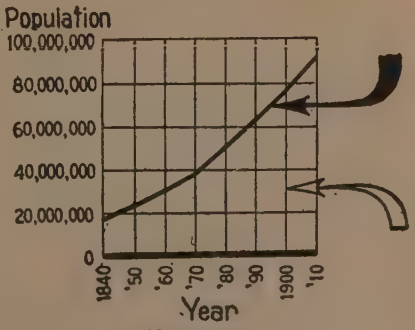


Illustration 10

11. In curves representing a series of observations, it is advisable, whenever possible, to indicate clearly on the diagram all the points representing the separate observations.

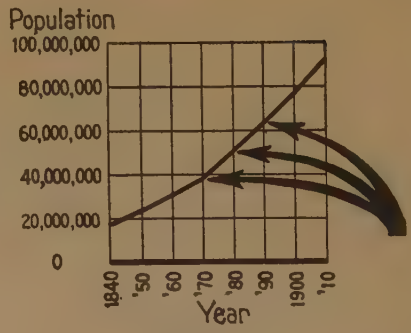


Illustration 11a

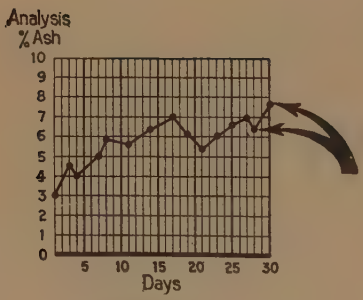


Illustration 11b

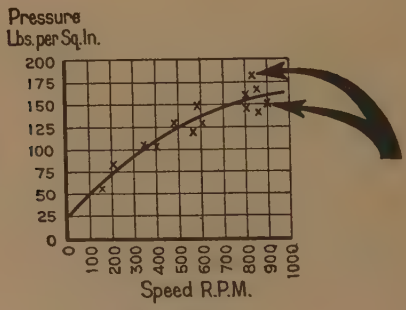


Illustration 11c

12. The horizontal scale for curves should usually read from left to right and the vertical scale from bottom to top.

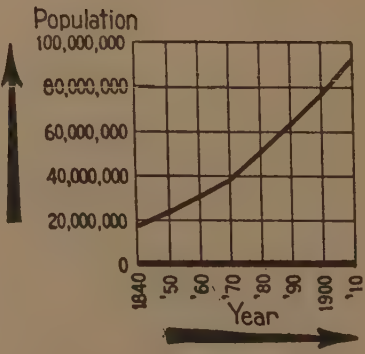


Illustration 12

13. Figures for the scales of a diagram should be placed at the left and at the bottom or along the respective axes.

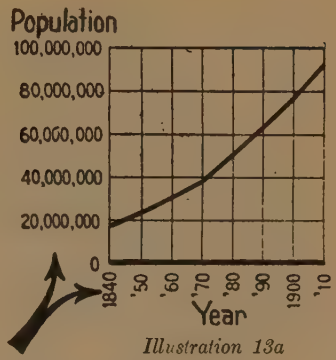


Illustration 13a

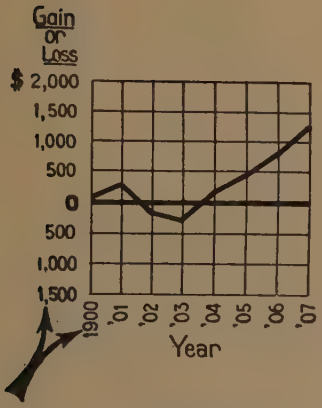


Illustration 13b

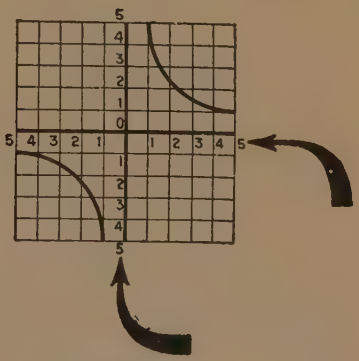


Illustration 13c

14. It is often desirable to include in the diagram the numerical data or formulæ represented.

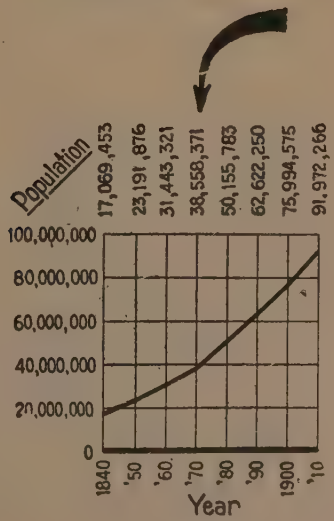


Illustration 14a

Illustration 14b

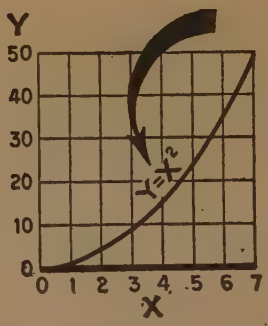
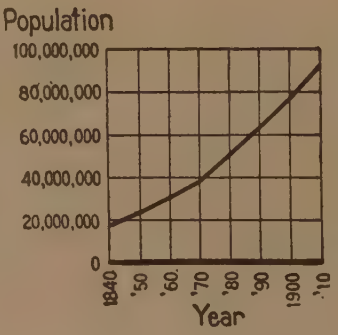


Illustration 14c

15. If numerical data are not included in the diagram it is desirable to give the data in tabular form accompanying the diagram.



Year	Population
1840	17,069,453
1850	23,191,876
1860	31,443,321
1870	38,558,371
1880	50,155,783
1890	62,622,250
1900	75,994,575
1910	91,972,266

Illustration 15

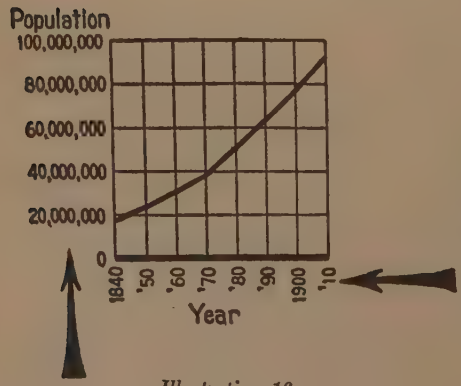
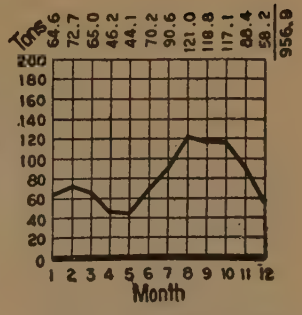


Illustration 16

16. All lettering and all figures on a diagram should be placed so as to be easily read from the base as the bottom, or from the right-hand edge of the diagram as the bottom.

17. The title of a diagram should be made as clear and complete as possible. Subtitles or descriptions should be added if necessary to insure clearness.



Aluminum Castings Output of Plant No. 2, by Months, 1914.

Output is given in short tons.

Sales of Scrap Aluminum are not included.

Illustration 17

REFERENCES

(1) *On Mathematical Concepts*

A discussion of the elementary mathematical concepts involved in graphic presentation and the employment of logarithms will be found in any of the standard textbooks on algebra and coördinate geometry. The treatment of these subjects in the books listed below will be found helpful to the student of statistics.

Griffin, F. L. *Introduction to Mathematical Analysis.*
Karsten, Karl G. *Charts and Graphs.*
Lipka, Joseph. *Graphical and Mechanical Computation.*
Mellor, J. W. *Higher Mathematics for Students of Chemistry and Physics.*
Schultze, Arthur. *Graphic Algebra.*
Steinmetz, C. P. *Engineering Mathematics.*
Whitehead, A. N. *An Introduction to Mathematics.*

(2) *On Graphic Methods*

Brinton, W. C. *Graphic Methods for Presenting Facts.*
Clark, Wallace. *The Gantt Chart.*

FIELD, J. H. *Some Advantages of the Logarithmic Scale in Statistical Diagrams.* Journal of Political Economy, October, 1917.

FISHER, IRVING. *The 'Ratio' Chart.* Quarterly Publications of the American Statistical Association, June, 1917.

HASKELL, A. C. *Graphic Charts in Business.*

HASKELL, A. C. *How to Make and Use Graphic Charts.*

KARSTEN, KARL. *Charts and Graphs.*

(The publishers and the dates of publication of the volumes named above are given in the bibliography at the end of this volume.)

CHAPTER III

THE ORGANIZATION OF STATISTICAL DATA: THE FREQUENCY DISTRIBUTION

The task of the statistician engaged in business or economic research includes the organization, analysis and interpretation of quantitative data relating to business affairs and to economic conditions. To these fundamental operations that of collecting the original data may be added, though more frequently data will be compiled directly from primary or secondary sources.

At the outset it is necessary to distinguish between the problems arising in the analysis of time series and those involved in the organization and analysis of materials in connection with which the time factor does not enter. In studying a time series the primary object is to measure and analyze the chronological variations in the value of the variable. Thus one may study variations in sales over a period of years, fluctuations in the production of bituminous coal, or changes in the general level of prices. Quite different is the procedure in the study of such a problem as income distribution at a given time. In this case we are desirous of knowing how many people in the United States fall in each of a number of income classes. The general problem of organization in this latter class of cases is to determine how many times each value of a variable is repeated and how these values are distributed. Data of this sort, when organized, constitute a *frequency series*, as opposed to the *time* or *historical series*. The methods appropriate to these two types of analysis differ funda-

mentally and will therefore be treated separately. In the present section we are concerned with the organization and preliminary analysis of data in connection with which the time element, while it may be present, does not enter as a factor.

Unorganized Data

When quantitative data of the type with which the statistician works are presented in a raw state they appear as unorganized masses of material, without form or structure. They may have been drawn from the production or sales records of a business establishment, or they may represent a miscellaneous collection of price quotations. If the data have been gathered by other agencies they may already have been arranged in the form of a general table, but this form may be entirely unsuited to the particular object in the mind of the investigator. The first task of the statistician is the organization of the figures in such a form that their significance, for the purpose in hand, may be appreciated, that comparison with masses of similar data may be facilitated, and that further analysis may be possible. Scientific method, it has been noted, involves *observation*, *inference* and *verification*. Data, the results of *observation*, must be put into definite form, must be given coherent structure, before the process of *inference* is possible.

The following figures, representing the earnings during a given week of 210 individuals engaged in piece work in a certain manufacturing establishment, will serve as an example of such data in their raw state:

Weekly Earnings of 210 Employees

$26.25	$28.70	$24.15	$29.75	$29.20	$30.60	$23.40	$24.75
26.70	24.35	25.75	27.20	28.30	25.25	27.75	27.60
28.20	27.30	27.80	26.35	27.40	28.30	26.60	25.75
27.70	28.60	25.30	27.80	26.40	27.30	28.35	27.00
24.30	27.80	27.60	26.30	27.40	23.50	29.60	27.80
27.60	25.35	27.55	29.00	24.10	27.00	24.50	27.25
26.15	29.30	23.10	27.10	28.50	27.45	26.15	28.35
27.95	25.55	27.55	26.60	24.25	30.00	28.55	28.00
27.30	27.90	25.25	24.10	27.45	24.55	26.55	27.55
26.75	31.00	24.00	25.35	26.50	28.30	27.95	25.55
30.25	28.55	26.75	24.60	25.75	26.55	27.80	28.90
29.55	30.00	24.60	25.75	26.30	27.00	28.25	25.25
25.75	26.25	26.30	26.75	27.90	28.30	25.70	26.30
26.60	27.00	30.75	28.60	28.10	23.50	24.75	25.15
26.30	27.25	28.15	29.10	30.10	29.90	28.55	27.30
26.55	27.55	23.00	24.50	22.55	26.55	27.55	28.10
30.70	28.60	27.90	26.80	24.10	25.25	26.30	27.90
26.90	25.30	25.80	28.85	27.55	27.30	25.00	26.00
26.55	27.80	28.60	30.55	29.50	24.10	25.15	27.15
28 10	26.30	27.10	24.60	27.80	26.30	27.90	29.80
24.10	25.15	27.50	24.25	25.70	26.80	30.15	29.30
28.15	28.65	24.55	25.85	26.10	27.00	26.80	27.55
29.00	23.00	28.60	29.30	28.55	28.80	27.55	23.60
26.10	27.15	25.75	26.80	27.15	26.30	28.55	25.80
24.55	25.80	26.75	27.30	27.55	28.25	25.60	26.30
26.85	27.30	28.10	32.00	28.15	26.30	27.75	26.25
28.60	26.00						

The Array

If these figures are arranged in order of magnitude something will have been done toward securing a coherent structure. The range covered and the general distribution throughout this range will then be clear, and the way will be prepared for further organization. When so arranged the following *array* is secured:

ARRAY: WEEKLY EARNINGS OF 210 EMPLOYEES

$22.55	$25.15	$26.15	$26.75	$27.45	$27.95	$28.60
23.00	25.15	26.15	26.75	27.45	27.95	28.65
23.00	25.15	26.25	26.80	27.50	28.00	28.70
23.10	25.25	26.25	26.80	27.55	28.10	28.80
23.40	25.25	26.25	26.80	27.55	28.10	28.85
23.50	25.25	26.30	26.80	27.55	28.10	28.90
23.50	25.25	26.30	26.85	27.55	28.10	29.00
23.60	25.30	26.30	26.90	27.55	28.15	29.00
24.00	25.30	26.30	27.00	27.55	28.15	29.10
24.10	25.35	26.30	27.00	27.55	28.15	29.20
24.10	25.35	26.30	27.00	27.55	28.20	29.30
24.10	25.55	26.30	27.00	27.55	28.25	29.30
24.10	25.55	26.30	27.00	27.60	28.25	29.30
24.10	25.60	26.30	27.10	27.60	28.30	29.50
24.15	25.70	26.30	27.10	27.60	28.30	29.55
24.25	25.70	26.30	27.15	27.70	28.30	29.60
24.25	25.75	26.35	27.15	27.75	28.30	29.75
24.30	25.75	26.40	27.15	27.75	28.35	29.80
24.35	25.75	26.50	27.20	27.80	28.35	29.90
24.50	25.75	26.55	27.25	27.80	28.50	30.00
24.50	25.75	26.55	27.25	27.80	28.55	30.00
24.55	25.75	26.55	27.30	27.80	28.55	30.10
24.55	25.80	26.55	27.30	27.80	28.55	30.15
24.55	25.80	26.55	27.30	27.80	28.55	30.25
24.60	25.80	26.60	27.30	27.80	28.55	30.55
24.60	25.85	26.60	27.30	27.90	28.60	30.60
24.60	26.00	26.60	27.30	27.90	28.60	30.70
24.75	26.00	26.70	27.30	27.90	28.60	30.75
24.75	26.10	26.75	27.40	27.90	28.60	31.00
25.00	26.10	26.75	27.40	27.90	28.60	32.00

FREQUENCY TABLES

While this array presents the figures in a shape much more suitable for study than the haphazard distribution first shown, there is still something to be desired before the mind can readily grasp the full significance of the data. The factory manager may see that the smallest amount earned during the week was $22.55, that the largest amount earned was $32.00, and that most of the employees earned between $25.00 and $29.00, but this is still a vague description of the data. By a process of grouping, that is, by

putting into common classes all individuals whose earnings fall within certain limits, a simplified and more compact presentation of the wage distribution may be obtained. The following table shows the results of this grouping process when the range of each class (the *class-interval*) is two dollars.

TABLE 6

Frequency Distribution of Employees

Classified on the Basis of Weekly Earnings (Class-interval = $2)

Weekly earnings	*Number earning stated amount (frequency)*
$22.00 to $23.99	8
24.00 to 25.99	48
26.00 to 27.99	96
28.00 to 29.99	47
30.00 to 31.99	10
32.00 to 33.99	1
	210

This table presents a condensed summary of the original figures, a summary which not only gives us the approximate range of the earnings, but shows, also, how the earnings of the 210 workers are distributed throughout this range. There has been a considerable loss of detail, it will be noted. From this table we may learn that there are 48 persons who earned during the given week between $24.00 and $25.99, but we cannot learn how the earnings of the 48 individuals were distributed throughout this range of two dollars. All may have earned exactly $24.00, so far as we may know from the figures shown in the table. This loss of detail is an inevitable accompaniment of the condensation and simplification which the process of classification involves.

If the size of the class-interval be decreased the loss of detail is less pronounced, though the increase in the number of classes means a more cumbersome table and one which presents a more complex picture to the eye. The tables which follow present the same data, classified with intervals of one dollar, fifty cents and twenty-five cents:

FREQUENCY DISTRIBUTIONS OF EMPLOYEES
Classified on the Basis of Weekly Earnings

TABLE 7 (Class-interval = $1)		TABLE 8 (Class-interval = 50 cents)		TABLE 9 (Class-interval = 25 cents)	
Weekly Earnings	*Frequency*	*Weekly Earnings*	*Frequency*	*Weekly Earnings*	*Frequency*
$22.00 to $22.99	1	$22.50 to $22.99	1	$22.75 to $22.99	1
23.00 to 23.99	7	23.00 to 23.49	4	23.00 to 23.24	3
24.00 to 24.99	21	23.50 to 23.99	3	23.25 to 23.49	1
25.00 to 25.99	27	24.00 to 24.49	11	23.50 to 23.74	3
26.00 to 26.99	42	24.50 to 24.99	10	23.75 to 23.99	0
27.00 to 27.99	54	25.00 to 25.49	12	24.00 to 24.24	7
28.00 to 28.99	34	25.50 to 25.99	15	24.25 to 24.49	4
29.00 to 29.99	13	26.00 to 26.49	22	24.50 to 24.74	8
30.00 to 30.99	9	26.50 to 26.99	20	24.75 to 24.99	2
31.00 to 31.99	1	27.00 to 27.49	24	25.00 to 25.24	4
32.00 to 32.99	1	27.50 to 27.99	30	25.25 to 25.49	8
	—	28.00 to 28.49	17	25.50 to 25.74	5
	210	28.50 to 28.99	17	25.75 to 25.99	10
		29.00 to 29.49	7	26.00 to 26.24	6
		29.50 to 29.99	6	26.25 to 26.49	16
		30.00 to 30.49	5	26.50 to 26.74	10
		30.50 to 30.99	4	26.75 to 26.99	10
		31.00 to 31.49	1	27.00 to 27.24	11
		31.50 to 31.99	0	27.25 to 27.49	13
		32.00 to 32.49	1	27.50 to 27.74	14
			—	27.75 to 27.99	16
			210	28.00 to 28.24	9
				28.25 to 28.49	8
				28.50 to 28.74	14
				28.75 to 28.99	3
				29.00 to 29.24	4
				29.25 to 29.49	3
				29.50 to 29.74	3
				29.75 to 29.99	3
				30.00 to 30.24	4
				30.25 to 30.49	1
				30.50 to 30.74	3
				30.75 to 30.99	1
				31.00 to 31.24	1
				31.25 to 31.49	0
				31.50 to 31.74	0
				31.75 to 31.99	0
				32.00 to 32.24	1
					—
					210

The four tables above (including Table 6) represent four different degrees of condensation of the same data. Tables 6, 7 and 8 present the same general characteristics: a small number of cases in the extreme classes and a more or less regular increase in the frequencies as the center of each of the distributions is approached. The departure from regularity becomes greater the greater the number of classes. Table 9, in which the class-interval is 25 cents, has 38 classes. In this table the distribution of cases throughout the range is highly irregular, with pronounced departures from symmetry. The structure of each of the other tables is orderly and approaches more closely a condition of symmetry. Each presents the wage data in condensed and compact form, so that one consulting the tables may learn of the size and distribution of weekly earnings in the factory in question much more readily than by reference to the chaotic collection of figures first shown. Such organized collections of data are termed *frequency distributions*, and their purpose, as the term implies, is to show in a condensed form the nature of the distribution of a variable quantity throughout the range covered by the values of the variable. The construction of such a table is the first step to be taken in the organization and analysis of quantitative data of the type represented above.

Steps in the Construction of a Frequency Table

This general introduction to the subject of frequency tables has left untouched many important matters in connection with their construction. It remains to present a summary statement of these details. It will be clear that the first step here taken, the arrangement of the items in order of magnitude, is unnecessary in the actual construction of such a table. Having determined the upper and lower limits through an inspection of the data, one has but to decide on the number of classes desired, write the

class-intervals on an appropriate blank sheet and proceed to tally the cases falling in each of the classes thus set off. When this process is completed the frequencies are computed and the totals arranged in tabular form of the type illustrated above. These simple operations involve decisions on a number of points, however.

Size of Class-Interval

In deciding upon the size of the class-interval (which is equivalent to deciding upon the number of classes) one fundamental consideration should be borne in mind, namely, that classes should be so arranged that there will be no material departure from an even distribution of cases within each class. This arrangement is necessary because of the fact that, in interpreting the frequency table and in subsequent calculations based upon it, the mid-value of each class is taken to represent the values of all cases falling in that class. Thus, in basing calculations upon Table 8, it is assumed that the 22 cases falling between $26.00 and $26.50 may all be represented by the mid-value of that class, $26.25. This assumption will seldom be strictly valid. In the case just cited reference to the original figures will show that it is not a correct assumption. Absolute accuracy would only be obtained by having a class for every value represented in the original figures. Since condensation is necessary an arrangement of classes should be secured which will minimize the error involved, without transgressing other requirements. Table 6 furnishes an example of class-intervals too wide for the material.

The requirement which has just been described clearly calls for a large number of classes. A second requirement, which ordinarily conflicts with this, is that the number of classes should be so determined that an orderly and regular sequence of frequencies is secured. If the classification is too narrow for the data regularity will not be attained in

this respect, and a table without structure or order will be secured. Table 9 fails to meet this requirement, as has been pointed out. It is desirable, also, that the number of classes be limited in order that the data may be easily manipulated and their significance readily grasped. As a general rule the class-intervals should be so adjusted that the number of classes is not less than 10 nor more than 25. The exact number which should be established in a given case will depend upon the nature of the data. In the above example Table 8, in which the class-interval is fifty cents, seems to conform most thoroughly to all these requirements.

Location of Class Limits

The location of class limits is a matter of considerable importance, for attention to this matter will simplify tabulation and facilitate later calculation. Tabulation of data is easiest when class limits are integers and the class-interval itself is a whole number. Calculation of averages and other statistical measures is facilitated when the mid-values of classes are integers. Suitable class limits and mid-points are usually secured when the data permit class-intervals of 5 or multiples of 5 to be employed, though such an arrangement is by no means essential.

Some types of data show a tendency to cluster or concentrate about certain values on the scale along which they are distributed. This is illustrated by the following figures which form part of a table showing the number of pieces of commercial paper discounted by the Federal Reserve Banks in 1921, distributed according to rates of discount or interest charged by member banks:

Rate (per cent)	*Number of pieces*
6	18,970
6¼	697
6½	4,616
6¾	135
7	17,362
7¼	10

Here is a quite obvious bunching about the integers, with a secondary concentration at each half of one per cent. No cases at all fall between the quarter values here shown. It is clear that in classifying such data the mid-points of the various classes should fall at those values about which the cases are concentrated, and class limits must be located with this end in view. For, as noted above, calculations based upon the frequency table are performed upon the assumption that all the items in each class are concentrated at the mid-point of that class. Thus, if a class interval of one half of one per cent were selected in the above example, the classes should extend from $5\frac{3}{4}$ to (but not including) $6\frac{1}{4}$, $6\frac{1}{4}$ to $6\frac{3}{4}$, etc., rather than from 6 to $6\frac{1}{2}$, $6\frac{1}{2}$ to 7, etc.

Accuracy of Observations and the Definition of Classes

In the construction of frequency tables it is essential that there be a clear definition of classes, so that there may be no uncertainty as to their range and no question as to the precise class in which a given case falls. A table with an arrangement similar to the following is sometimes encountered:

Class-interval	*Frequency*
0 to 10	3
10 to 20	8
20 to 30	15
30 to 40	6
40 to 50	2

In the absence of explanation, a question arises at once as to whether a case with a value of 10 would fall in the first or in the second class. It is highly desirable that the range of each class be indicated in some such way as the following, in order that this ambiguity may not arise:

Class-interval	*Frequency*
0 to 9.9	3
10 to 19.9	8
20 to 29.9	15
30 to 39.9	6
40 to 49.9	2

This procedure solves the difficulty, however, only in case the observations are accurate to the nearest tenth. If the observations are accurate only to the nearest unit (that is, if the cases given a value of 10 actually lie between 9.5 and 10.5) a mere change in the description of the class-range does not solve the problem of allocating a case at the class limit. In such a case an observation falling at a class-boundary may be cut in two, one half being allocated to each of the adjacent classes.

Yule [1] lays down the useful principle that in fixing a class boundary the limit should be carried to a farther place in decimals, or a smaller fraction, than the values of the individual cases as originally recorded. Thus, in the preceding example, if observations were correct to the nearest tenth, it would mean that a value recorded as 9.9 actually lay between 9.85 and 9.95. In accurately describing the classes, therefore, the intervals should be given as 0 to 9.95, 9.95 to 19.95, etc. It should be noted that the values of the mid-points, with these class limits, would be 4.95, 14.95 etc. In presenting and using the table as given above the real meaning of the class limits should be borne in mind. In all cases class boundaries must be fixed with reference to the accuracy of the observations.

The work of tabulation is simplified if, in designating a class, both limits are stated, as above. Errors are likely if only the lower limit of each class is given, or if the mid-point alone is designated. It is desirable, however, particularly if calculations are to be based upon the table, to include a separate column showing the values of the mid-points of the various classes.

Other Requirements

Class-intervals should be uniform throughout the table in order that all classes may be comparable. Occasionally

[1] *An Introduction to the Theory of Statistics*, 81.

tables are published with varying class-intervals, so that on one section of the scale the number of items falling within a class having an interval of 5 is given, and on another section of the scale the number of items falling within a class having a range of 10 is given. Obviously, comparison of classes is impossible. It may be desirable to show in more detail the cases falling within certain ranges on the scale, but this end is best achieved by the construction of a supplementary table relating only to the cases falling within this restricted section. The utility of the main table is not lessened thereby.

Similar in nature is the requirement that there should be no indeterminate classes, that is, classes the ranges of which are not defined. Had all the individuals making $30.00 and over in the illustration of piece-work earnings been entered in a class with the designation "$30.00 and over," the upper limit of this class would have been quite uncertain. This fault in a table is a vital one when it is desired to base calculations upon the data contained in the table. When there are several extreme cases the inclusion of such classes is sometimes unavoidable, but when this is done the actual values of the cases included in such "open end" classes should be given in a footnote to the table.

The errors described in the two preceding paragraphs are exemplified in the table below.

TABLE 10

Frequency Distribution of Employees, ABC Company Classified on the Basis of Daily Wages

Wage group	*Number of workers in each class (frequency)*
Less than $3.00	15
$3.00 to 3.99	30
4.00 to 4.99	40
5.00 to 6.49	30
6.50 to 8.99	10
9.00 and over	5
	130

In this case the ranges of the two "open-end" classes are not known. The ranges of the intermediate classes vary, being $1.00 for two classes, $1.50 for one class, and $2.50 for one class. Such a table is of little value.

The Structure of Statistical Tables

The preceding discussion has been confined to certain more or less technical problems which arise in the construction of a frequency table. Nothing has been said directly as to the form of the completed table, the arrangement of columns and rows, the title, the notation. No general principles of tabular arrangement have been laid down. While no detailed treatment of these principles is possible within the scope of the present discussion, certain general considerations relating to the structure of statistical tables may be suggested.

The statistical table is merely a device for presenting in summary fashion a mass of quantitative data. Unless the summary be clear, significant, concise and readily interpreted nothing has been gained by the process of tabulation and classification. A sprawling, formless table is like a rambling, unintelligible discourse. There must be a purpose in back of each table, and this purpose should be clearly brought out in its arrangement. The means by which this purpose may be attained in a given case must be determined with reference to the specific conditions affecting that case, but standard practices should be followed, in so far as possible. The following general principles will be found helpful in deciding upon the form and arrangement of statistical tables:

1. The title should constitute a clear, concise and complete description of the material assembled in the table.
2. Headings of columns and rows should be concise and unambiguous.

3. Variable quantities should increase from left to right and from top to bottom, when such arrangement is feasible.
4. Columns and rows may be numbered to facilitate reference to the table.
5. The units of measurement employed should be clearly indicated.
6. Sources should be given in all cases.
7. The table should constitute a unit, self-sufficient and self-explanatory. All explanations necessary for its interpretation should be included as integral parts of the table, or in the form of footnotes.

Graphic Representation of Frequency Distributions

Frequency distributions of the type illustrated above serve a very important statistical function in presenting a

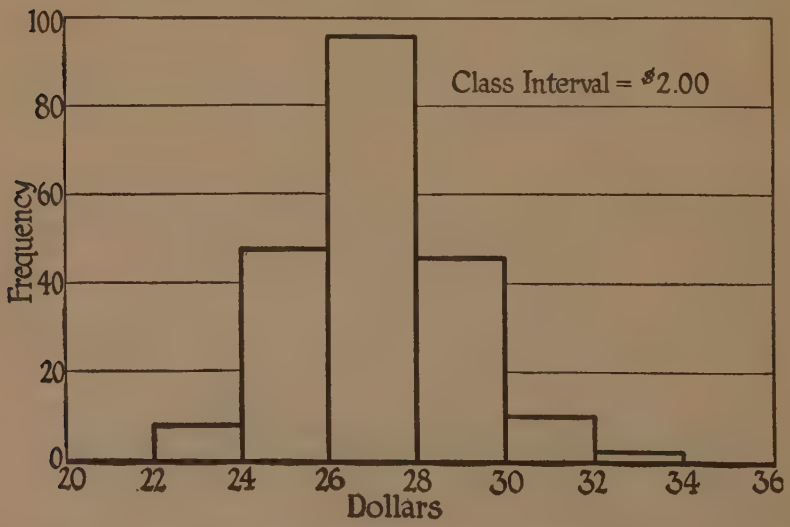


Fig. 22. — Column Diagram: Distribution of 210 Employees Classified on the Basis of Weekly Earnings. (Class-interval = $2.00)

compact summary of data, and in preparing these data for further manipulation. Such distributions may be presented not only in tabular form, but graphically, utilizing

the general principles of the coördinate system which were explained above. Many of the characteristic features of a frequency distribution are most clearly revealed when the graphic method is adopted.

Table 6, presenting the weekly earnings of 210 employees, with a class-interval of two dollars, is depicted

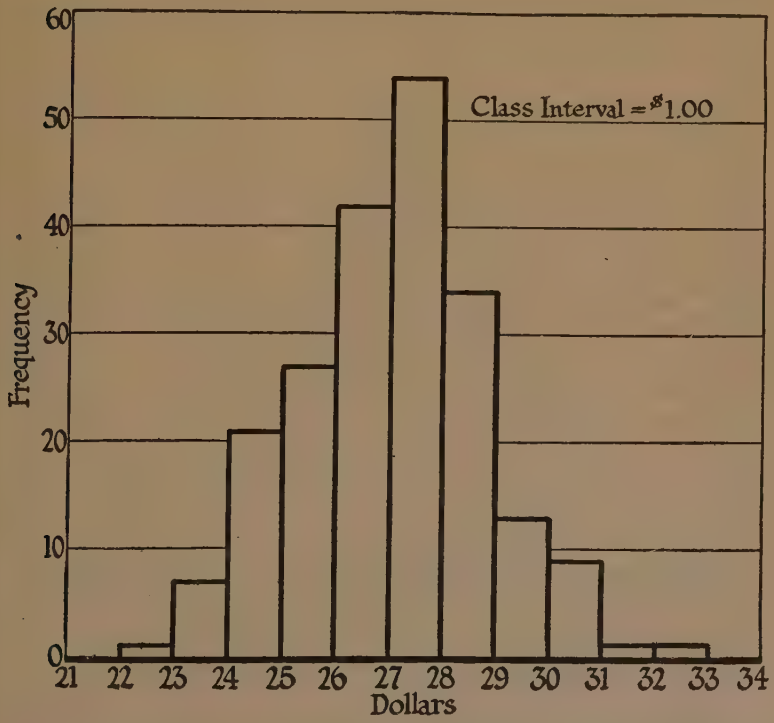


FIG. 23. — Column Diagram: Distribution of 210 Employees Classified on the Basis of Weekly Earnings. (Class-interval = $1.00)

graphically in Fig. 22. In this figure class-intervals are plotted along the x-axis and the corresponding class-frequencies along the y-axis, appropriate scales being selected. The fact should be noted that the scale of abscissas starts not with zero, but with $20. For convenience in presentation that part of the scale extending from 0 to $20 is omitted.

The student should bear this in mind in seeking to secure a correct impression of the relations between the two variables plotted. In constructing such a figure, which is termed a *column diagram* or *histogram*, short horizontal lines are drawn connecting the points plotted to represent the upper and lower limits of each class-interval. In inter-

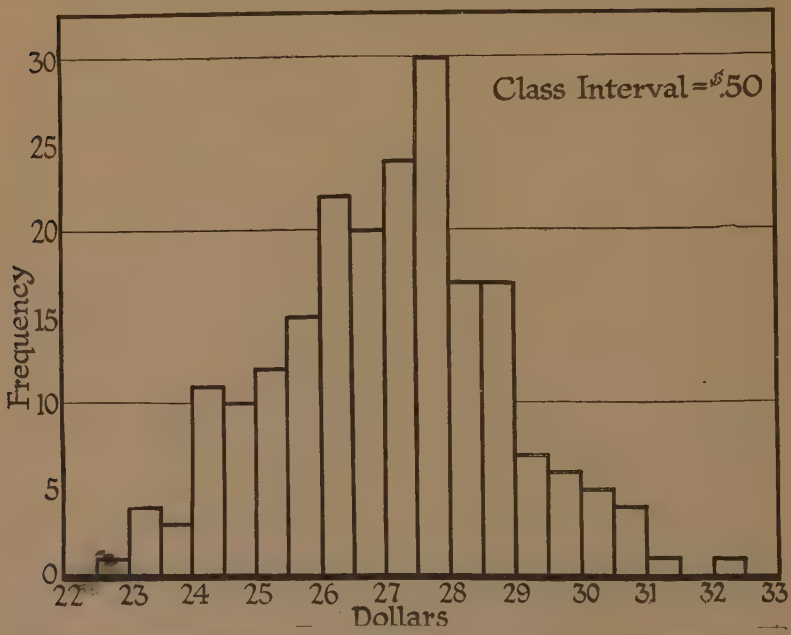


FIG. 24. — Column Diagram: Distribution of 210 Employees Classified on the Basis of Weekly Earnings. (Class-interval = $.50)

preting this diagram it should be noted that the areas of the different rectangles are proportional to the number of cases represented, the total area representing the entire 210 cases. This device thus presents to the eye a very clear picture of the distribution, showing quite unmistakably the relative number of workers falling in each of the wage classes.

The classes in this case are so large, however, that some violence is done to the facts. So many details are lost

that a true conception of the disposition of the items is not given. Fig. 23 is a histogram depicting the distribution of cases when a class-interval of one dollar is used. In this case, with smaller steps, we approach more closely an orderly and symmetrical distribution. The same is true of Fig. 24 which shows the distribution when the class-interval is

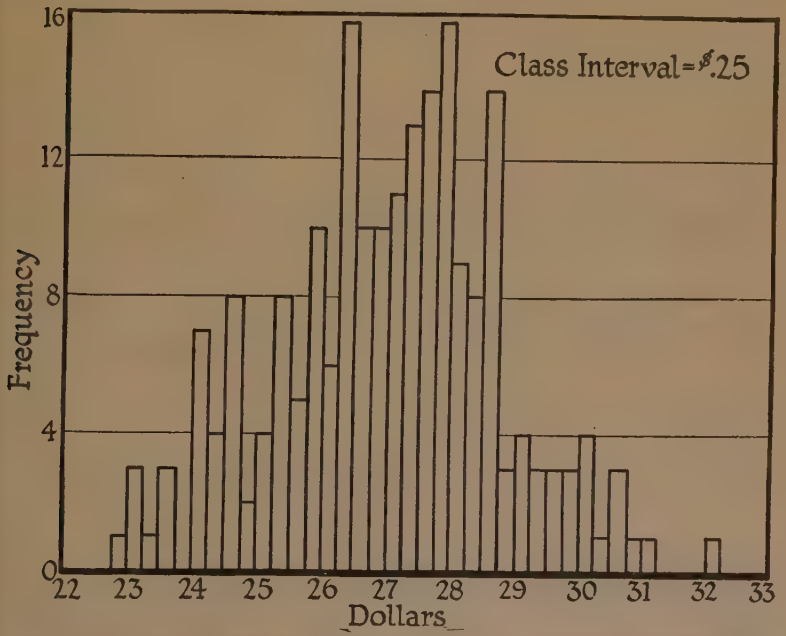


FIG. 25. — Column Diagram: Distribution of 210 Employees Classified on the Basis of Weekly Earnings. (Class-interval = $.25)

fifty cents. The distribution represented in Fig. 25 has a class-interval of twenty-five cents which, as has been pointed out, is too narrow for the data, with the result that a quite irregular structure is secured. (It should be noted that the vertical scale is not the same in these four figures, so that comparison with respect to class-frequencies is only possible by reference to the scale figures.)

Frequency polygons corresponding to the histograms of Figs. 22, 23, and 24 are shown in Figs. 26, 27, and 28.

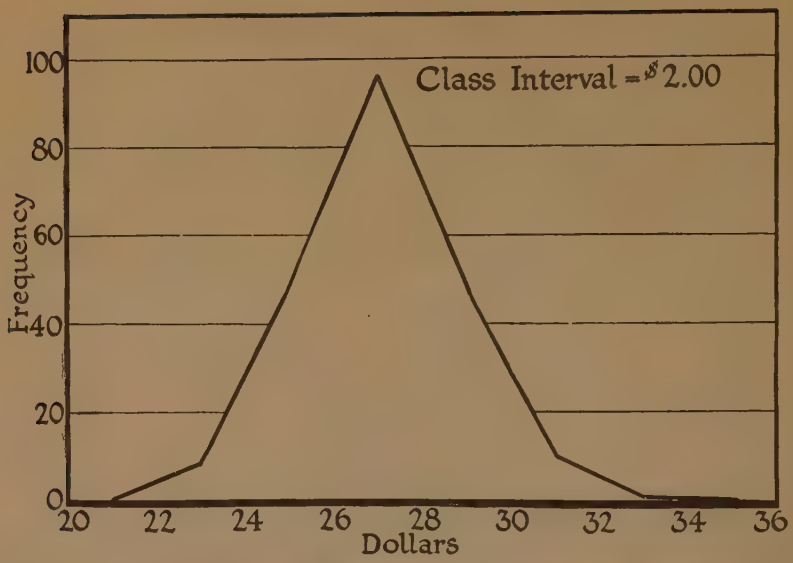


FIG. 26. — Frequency Polygon: Distribution of 210 Employees Classified on the Basis of Weekly Earnings. (Class-interval = $2.00)

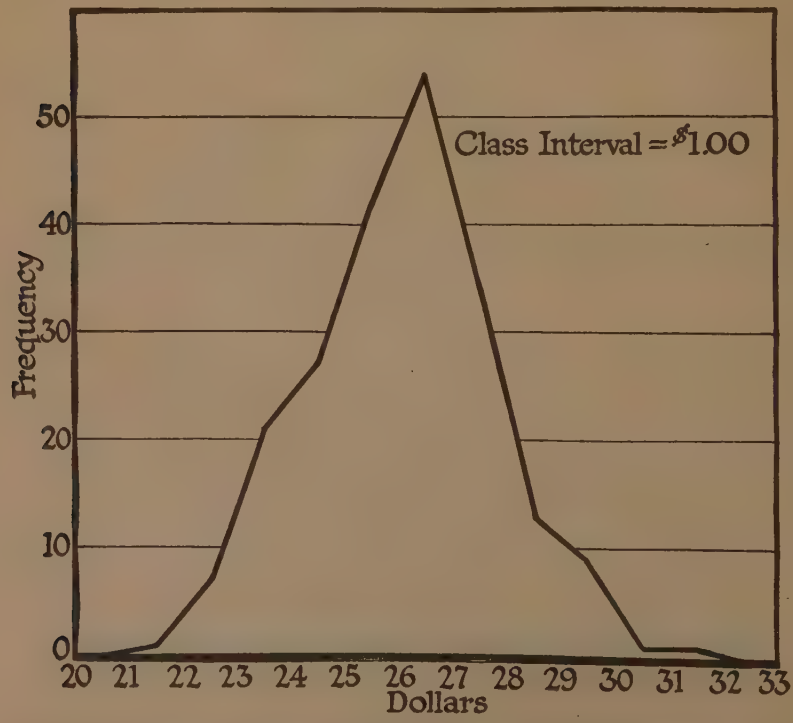


FIG. 27. — Frequency Polygon: Distribution of 210 Employees Classified on the Basis of Weekly Earnings. (Class-interval = $1.00)

Each of these polygons has been constructed by plotting as abscissas the mid-points of the class-intervals, and as ordinates the class-frequencies, the points thus secured being connected by a broken line. In completing such a figure

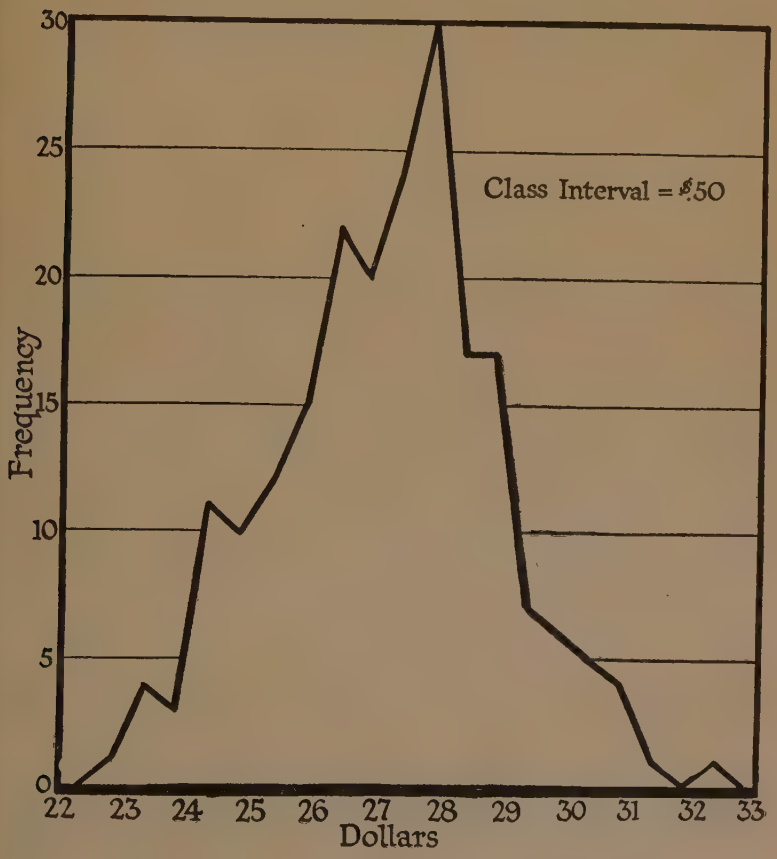


FIG. 28. — Frequency Polygon: Distribution of 210 Employees Classified on the Basis of Weekly Earnings. (Class-interval = $.50)

the class next below the lowest one on the scale and the class next above the highest one on the scale are included, the class-frequency being zero in each case. The ends of the polygon thus connect with the base line at the mid-

points of these two extra classes. In the case of the frequency polygon the entire area under the curve represents the entire number of cases, but the area of a given interval cannot be taken to be proportional to the number of cases in that interval, because of irregularities in the distribution on either side of the given class. The heights of the ordinates at the mid-points of the various classes are, of course, scaled to represent the class-frequencies.

The Smoothing of Curves

Attention is again called to the results secured with varying class-intervals. As the class-interval is decreased, up to a certain point, the histograms and polygons become smoother and more regular. Beyond that point breaks begin to appear in the data; the regular change in class-frequencies which was found when the classes were larger is broken by the appearance of irregular classes which seem to depart from the general rule. In Fig. 25 these have become quite pronounced. Such irregularities, it is obvious, are exceptions to a general rule which seems to prevail, the general rule that the numbers of workers falling within the different wage classes increase from the lower limit of earnings up to a maximum in the neighborhood of $27.50, and then decrease till, at the upper limit of $32, but one worker is found. Since all the 210 individuals are engaged in the same work, and since their earnings depend only upon their rapidity and skill, one would expect a quite regular increase and decrease. If we had figures not for one week only, but for 52 weeks, and took the average weekly earnings of each of the 210 workers for the year, we should expect greater regularity with the smaller class-intervals than is actually found, since the accidental fluctuations peculiar to one week alone would thus be eliminated. Or, if we had earnings during one week for 10,920 workers (52 times 210), the same result would be secured.

Thus, if regularity and smoothness are to be secured, it is essential not only to decrease the size of the classes but also to increase the number of cases, in order that the accidental irregularities which affect a small number of observations may be eliminated. A refined classification with a small number of cases leads to the condition exemplified in Fig. 25. But such an increase in the number of cases is, in general, a practical impossibility. We wish, if possible, to develop a feasible method of approximating the distribution which would be secured with very small class-intervals and a very large number of cases. Such an approximation is possible through the device of curve-smoothing. By this method we may secure a smooth *frequency curve* which lacks the irregularities occasioned by minor fluctuations.

Such a smooth frequency curve serves to represent the true underlying distribution of the data. It was pointed out that areas in the frequency polygon are not proportional to the number of cases included, the cause lying in the irregularities of the data. In a smoothed frequency curve these irregularities have been eliminated, and the area between ordinates erected at given points on the scale of abscissas is assumed to be proportional to the theoretical frequency of cases between the given values. Moreover, a smooth trend having been established, frequencies for intermediate values not shown in the original table may be determined by interpolation.[1]

The following data,[2] representing the distribution in 1918 of personal incomes below $4000, will serve to exemplify the smoothing process.

[1] The limitations of practical statistical work are such that there must of necessity be many gaps in the data. The given values of the variables are not continuous. Interpolation is the process of estimating values of a variable quantity between given values, or of locating a point on a curve between given points. That interpolation is most accurate which leads to estimated values having the highest degree of consistency with the given values.

[2] From Vol. I, *Income in the United States*, National Bureau of Economic Research. New York, Harcourt, Brace & Co., 1921, 132–33.

TABLE 11

Distribution of Income Among Personal Income Recipients in 1918

(Including all personal incomes below $4000)

Income Class [1]	*Number of Persons* [2]
$ 0 to $ 100	62,809
100 to 200	103,704
200 to 300	209,087
300 to 400	489,963
400 to 500	961,991
500 to 600	1,549,974
600 to 700	2,154,474
700 to 800	2,668,466
800 to 900	3,013,034
900 to 1,000	3,144,722
1,000 to 1,100	3,074,351
1,100 to 1,200	2,850,526
1,200 to 1,300	2,535,285
1,300 to 1,400	2,205,728
1,400 to 1,500	1,832,230
1,500 to 1,600	1,512,649
1,600 to 1,700	1,234,397
1,700 to 1,800	999,996
1,800 to 1,900	811,236
1,900 to 2,000	663,789
2,000 to 2,100	549,787
2,100 to 2,200	463,222
2,200 to 2,300	395,115
2,300 to 2,400	340,141
2,400 to 2,500	295,490
2,500 to 2,600	258,650
2,600 to 2,700	227,731
2,700 to 2,800	201,488
2,800 to 2,900	178,901
2,900 to 3,000	154,499
3,000 to 3,100	142,802
3,100 to 3,200	128,217
3,200 to 3,300	115,583
3,300 to 3,400	104,504
3,400 to 3,500	94,803
3,500 to 3,600	86,405
3,600 to 3,700	79,023
3,700 to 3,800	72,562
3,800 to 3,900	66,900
3,900 to 4,000	61,894

[1] The definition of classes used is equivalent to "$0 to and not including $100," etc. Thus an individual with an income of $100 would fall in the second class.

[2] The Bureau's report states "The numbers below are given to the nearest unit. It is not pretended that such arithmetic accuracy is anything more than technical."

Figures 29, 30, and 31 present column diagrams of these income data, grouped with class intervals of $500, $200 and $100. As the class-interval is decreased the histograms become more regular and uniform, but our original data permit us to carry this process only to the point where the class-interval is $100. Our problem is to determine the underlying distribution which the data approximate more and more closely as the class-interval is lessened. If

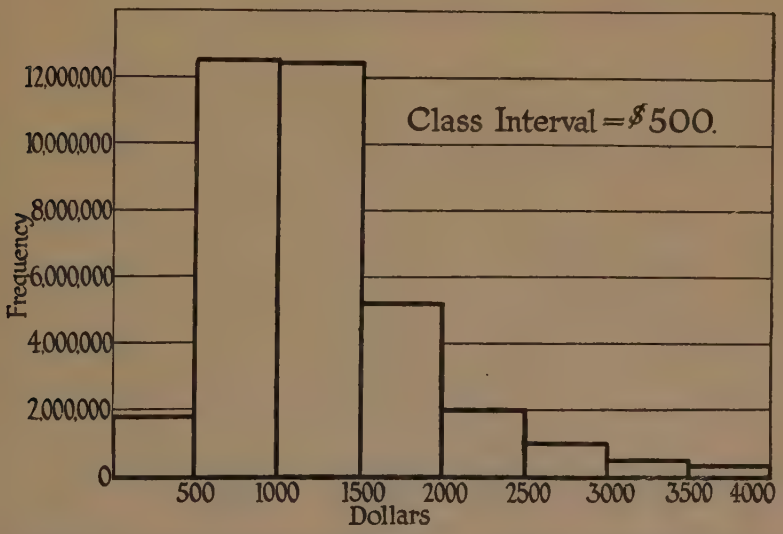


FIG. 29. — Column Diagram: Distribution of Personal Income Recipients in the United States, 1918. Including all Recipients of Incomes below $4,000 (Class-interval = $500)

we replace the broken line of the histogram by a smooth curve *enclosing the same total area* as the histogram and so drawn through the points of the histogram that *the area cut from each rectangle is approximately equal to the area added to the same rectangle by the curve*, we will have a frequency curve representing the desired distribution. The requirement that the same total area be enclosed is fundamental. Exceptions to the rule concerning the area of individual rectangles will frequently occur because of the

existence of quite irregular classes, but as a general working principle it is helpful. (More refined methods of fitting a smooth curve to data will be discussed at a later point, but a process of smoothing by inspection such as that described above gives a fairly close approximation to the required curve.)

Figure 32 illustrates the result of smoothing the histogram of income distribution shown in Fig. 31. Here the

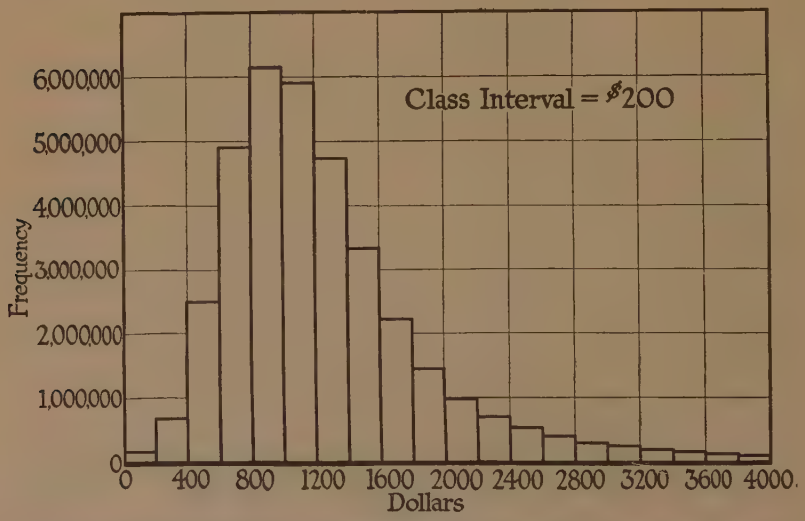


FIG. 30. — Column Diagram: Distribution of Personal Income Recipients in the United States, 1918. Including all Recipients of Incomes below $4,000 (Class-interval = $200)

quite artificial jumps between income classes are smoothed out, and we secure the graduation by infinitesimal increments which we should expect to find when the incomes of so many millions of persons are included. Here we have that which we desired — an approximation to the true underlying distribution, with the sharp breaks resulting from the method of classification eliminated.

Continuous and Discrete Series

The logical validity of the smoothing process is dependent upon the nature of the data being manipulated. From this point of view frequency series of the type discussed above may be divided into two classes, *continuous series* and *non-continuous* or *discrete series*. A continuous series is one in which the values of the independent variable

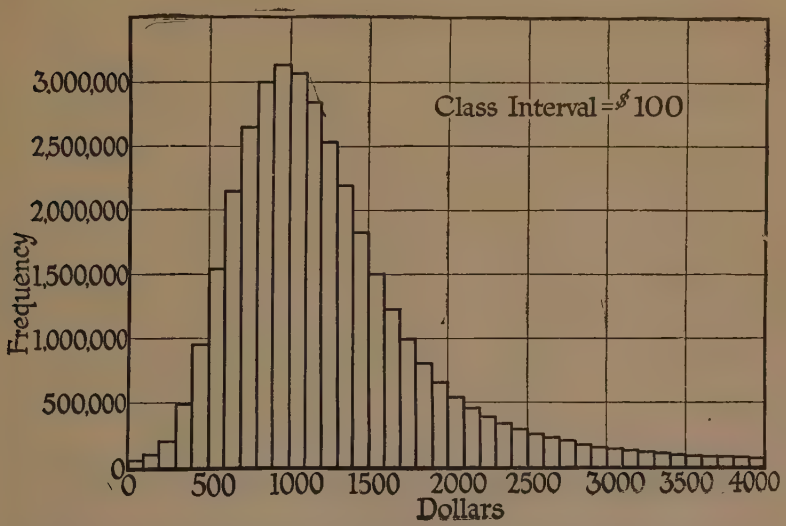


Fig. 31. — Column Diagram: Distribution of Personal Income Recipients in the United States, 1918. Including all Recipients of Incomes below $4,000 (Class-interval = $100)

increase or decrease by increments which are infinitely small. A discrete series is one in which the phenomena represented by the independent variable always change in value by definite amounts. The curve of underlying values rises not smoothly, as for the continuous series, but by jumps.

The fact should be emphasized that in making this distinction we are speaking of the values as they would be found in the underlying universe of phenomena from which

the actual bodies of material we study are drawn. Any given sample, whether representing continuous or discrete series, will be marked by breaks in the values of the independent variable. This will be true, in the case of a continuous series, because of the limitations upon the instruments and senses we use in measuring. Thus if the heights of 100 men be measured, the independent variable of the frequency series (height) will increase by finite amounts.

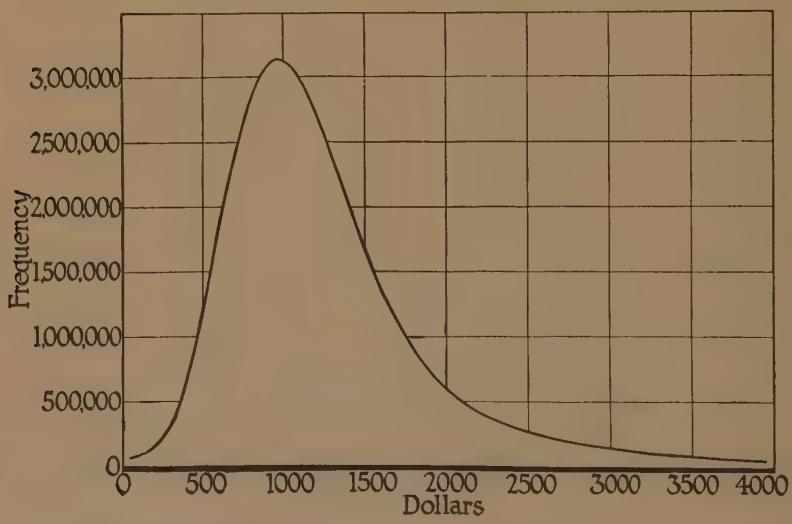


FIG. 32. — Frequency Curve: Distribution of Personal Income Recipients in the United States, 1918. Including all Recipients of Incomes below $4,000 (Derived from the column diagram with class-interval of $100)

We may measure to the nearest inch, or perhaps to the nearest eighth of an inch. Yet if ten thousand or ten million men were arranged in order of height the differences between successive individuals would be infinitely small. Height is a continuous variable, even though the values found in a given sample are marked by discontinuity.

Quite different is the distribution of such a variable as interest or discount rates. If one were to secure 100 such quotations and rank them in the order of size the varia-

tions would be discontinuous, as in the sample of men whose heights were measured. But in the case of heights the underlying values, if they could be determined for a large population, would be marked by continuous variation, whereas, were an infinite number of discount rate quotations secured, there would still be breaks in the sequence. Discount rates increase or decrease by one quarter or one half of one per cent, not by infinitesimal amounts. Such a series is termed discrete, or non-continuous.

The smoothing process provides a means of securing an approximation to the distribution of values as they would be found if a sample could be increased indefinitely in size. It is based upon the assumption that the irregularities found in the sample actually studied are accidental, and that the underlying values would show continuous and unbroken variation. Obviously, therefore, it is only fully justified when applied to a continuous series. A histogram of human heights may be smoothed in order to secure a representation of the true underlying distribution in the population at large, and interpolation based upon this smoothing process is valid. But smoothing is quite illogical for a markedly discontinuous series. It would be meaningless to construct a smooth curve showing the distribution of discount rates for the purpose of securing the theoretical frequency of a rate of 4.3675 per cent. In practical statistical work, however, it is frequently helpful to handle discrete series as though they were continuous, and in these cases the smoothing device may be employed. But in the interpretation and use of the smoothed curve the important logical distinction between continuous and discontinuous variation should be kept clearly in mind.[1]

[1] For a fuller discussion of the points discussed in this section see G. H. Knibbs, "The Theory and Justification of Curve Smoothing," in H. Secrist, *Readings and Problems in Statistical Methods*. N. Y., Macmillan, 1920, 278–82.

Cumulative Arrangement of Statistical Data

For certain purposes it is desirable to arrange data cumulatively, rather than in separate and exclusive classes of the type illustrated in the frequency tables presented above. The following material will illustrate some of the advantages of this arrangement.

In a study of the durability of telephone poles [1] these results were secured:

Table 12

Frequency Distribution of 248,707 Telephone Poles, Classified According to Length of Life

Length of Life (*years*)	*Number of Poles* (*frequency*)
0– 0.9	1,150
1– 1.9	4,221
2– 2.9	10,692
3– 3.9	13,966
4– 4.9	16,633
5– 5.9	18,211
6– 6.9	19,011
7– 7.9	19,260
8– 8.9	20,909
9– 9.9	19,879
10–10.9	20,764
11–11.9	15,454
12–12.9	14,237
13–13.9	13,779
14–14.9	9,764
15–15.9	8,534
16–16.9	7,659
17–17.9	6,918
18–18.9	4,591
19–19.9	1,798
20–20.9	815
21–21.9	313
22–22.9	102
23–23.9	47

The table shows that 1,150 poles were scrapped during the first year of use, that 4,221 were scrapped after reaching

[1] "Replacement Insurance," Edwin Kurtz. *Administration*, July, 1921, 41–69.

the age of one year and before reaching the age of two years, and so on. This is simply a frequency table of the ordinary type. A much more significant arrangement for many purposes is secured when the figures are assembled cumulatively, as in the following table.

TABLE 13

Cumulative Distribution of 248,707 Telephone Poles, Classified According to Length of Life

(Cumulated upward)

Length of Life	*Number of Poles Surviving (frequency)*
Less than 1 year	1,150
" " 2 years	5,371
" " 3 "	16,063
" " 4 "	30,029
" " 5 "	46,662
" " 6 "	64,873
" " 7 "	83,884
" " 8 "	103,144
" " 9 "	124,053
" " 10 "	143,932
" " 11 "	164,696
" " 12 "	180,150
" " 13 "	194,387
" " 14 "	208,166
" " 15 "	217,930
" " 16 "	226,464
" " 17 "	234,123
" " 18 "	241,041
" " 19 "	245,632
" " 20 "	247,430
" " 21 "	248,245
" " 22 "	248,558
" " 23 "	248,660
" " 24 "	248,707

It is important to note that it is possible to cumulate a frequency series in two different ways. From the above table we may determine readily the number failing to attain any given age. It is often more convenient to reverse the process, so that the table will enable the total

number above any given value to be immediately determined. When the telephone pole figures are thus *cumulated downward* the following table is secured.

TABLE 14

Cumulative Distribution of 248,707 Telephone Poles, Classified According to Length of Life

(Cumulated downward)

(1) *Length of Life*	(2) *Number of Poles Surviving (frequency)*	(3) *Per cent*
0 and more	248,707	100.0
1 year " "	247,557	99.5
2 years " "	243,336	97.8
3 " " "	232,644	93.6
4 " " "	218,678	88.0
5 " " "	202,045	81.2
6 " " "	183,834	73.8
7 " " "	164,823	66.3
8 " " "	145,563	58.5
9 " " "	124,654	50.1
10 " " "	104,775	42.1
11 " " "	84,011	33.8
12 " " "	68,557	27.6
13 " " "	54,320	21.8
14 " " "	40,541	16.3
15 " " "	30,777	12.4
16 " " "	22,243	8.9
17 " " "	14,584	5.9
18 " " "	7,666	3.1
19 " " "	3,075	1.2
20 " " "	1,277	0.5
21 " " "	462	0.2
22 " " "	149	0.06
23 " " "	47	0.02
24 " " "	0	0.00

Cumulative tables such as those given above have distinct advantages in handling many types of data. Life tables are generally presented in this form. The scientific study of depreciation will lead to the construction of elaborate "mortality tables" for various types of equipment, and these will be most useful in the cumulative

form. It is frequently desirable to reduce the frequencies to percentages, as in column (3) of Table 14, though it should not be forgotten that the significance of the percentages depends upon the absolute numbers upon which they are based.

The Ogive, or Cumulative Frequency Curve

The general utility of such cumulated data is limited by the classification system necessarily adopted in condensing

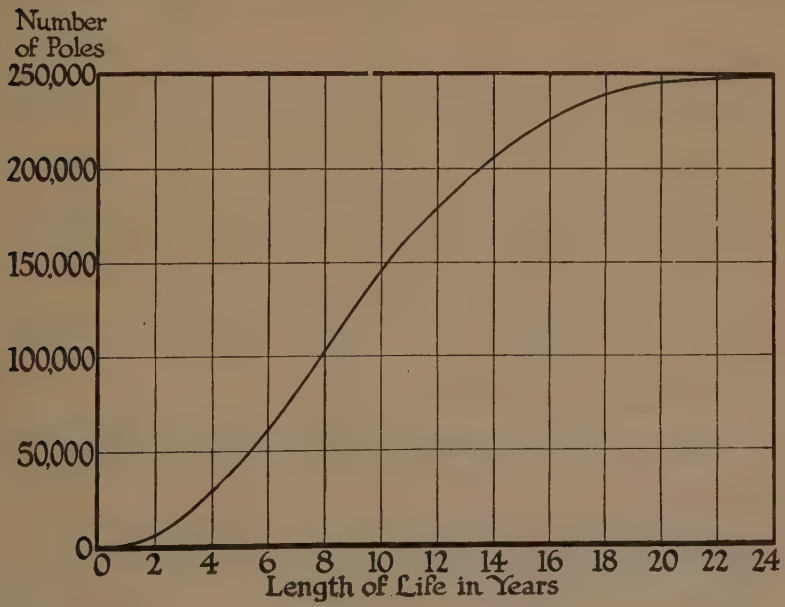


Fig. 33. — Cumulative Frequency Curve: Distribution of Telephone Poles Classified according to Length of Life. (Cumulated upward)

the material. Unless we interpolate mathematically we are limited to the points on the scale actually noted in the two tables. For this reason, a generalized cumulative curve similar to the smoothed frequency curve described in the preceding section is desirable. If the values given in Table 13 be plotted on coördinate paper (the length of life

in each case as abscissa, and the corresponding number of poles as ordinate) and a smooth curve drawn through the points thus plotted, the cumulative frequency curve shown in Fig. 33 is secured. In Fig. 34 the data of Table 14 are plotted.

Such a curve constitutes one of the most effective and useful representations of a frequency series. It is obvious

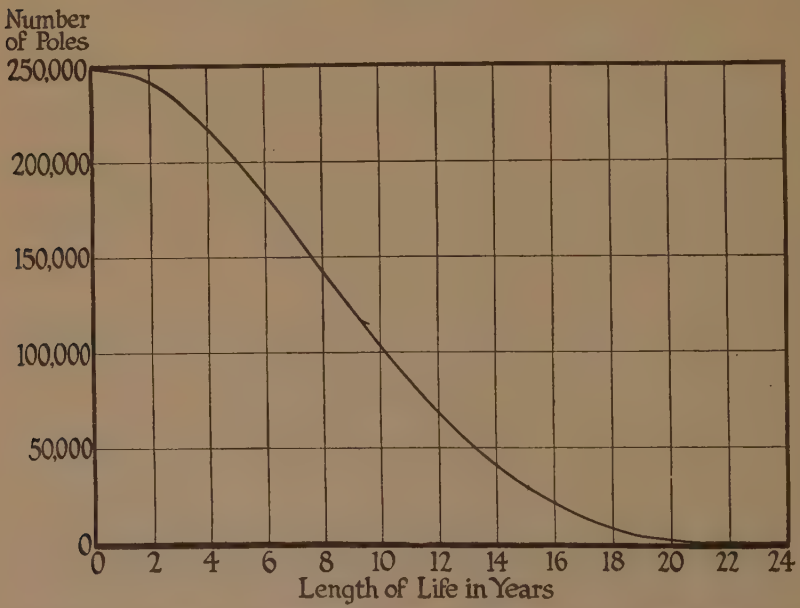


FIG. 34. — Cumulative Frequency Curve: Distribution of Telephone Poles Classified according to Length of Life. (Cumulated downward)

that the limitations of the particular class-interval adopted are in large part removed; the shape of the curve will be fundamentally the same, though the class-interval and number of classes may vary. Frequency curves of the usual type may not be compared unless the groupings are the same, but cumulative frequency curves are subject to no such restriction. Moreover, uneven class-intervals do not distort the *ogive*, or cumulative curve, as they do the ordinary frequency curve.

The cumulative curve is particularly well adapted to interpolation. Thus if it is desired to know the number of poles surviving less than $15\frac{1}{2}$ years, the value of the ordinate of the curve having $15\frac{1}{2}$ as abscissa may be approximated from Fig. 33. A value of 222,000 is secured. If the number surviving $8\frac{1}{2}$ years or more is desired, a similar estimate may be made from Fig. 34. The interpolated figure in this case is 135,000.

Another type of interpolation possible with such a curve is the determination of the number of cases falling within any given interval. One is not limited to the class-intervals marked out in the original tables. For instance, it may be desirable to know the number of poles surviving more than $10\frac{1}{2}$ but less than 15 years. Reading from the table or from the chart we find that 217,930 poles survived less than 15 years. Interpolating on the chart in the manner described above a figure of 154,000 is secured for the number surviving less than $10\frac{1}{2}$ years. Subtracting the latter figure from the former we have 63,930 as the number of poles falling within the $10\frac{1}{2}$ to 15 years interval. The figure is, of course, an approximation to the true value, as are all values secured through such smoothing and interpolation.

It should be noted that the ogive may be derived directly from the array, without the formation of a frequency table as an intermediate step. This curve, in fact, may be looked upon as merely a graphic representation of the array. It represents one of the simplest forms of statistical organization, as well as one of the most effective methods of manipulating quantitative data.

Relation Between the Ogive and the Frequency Curve

The ogive and the frequency curve are merely two different arrangements of precisely the same material, each arrangement having certain distinctive advantages. The

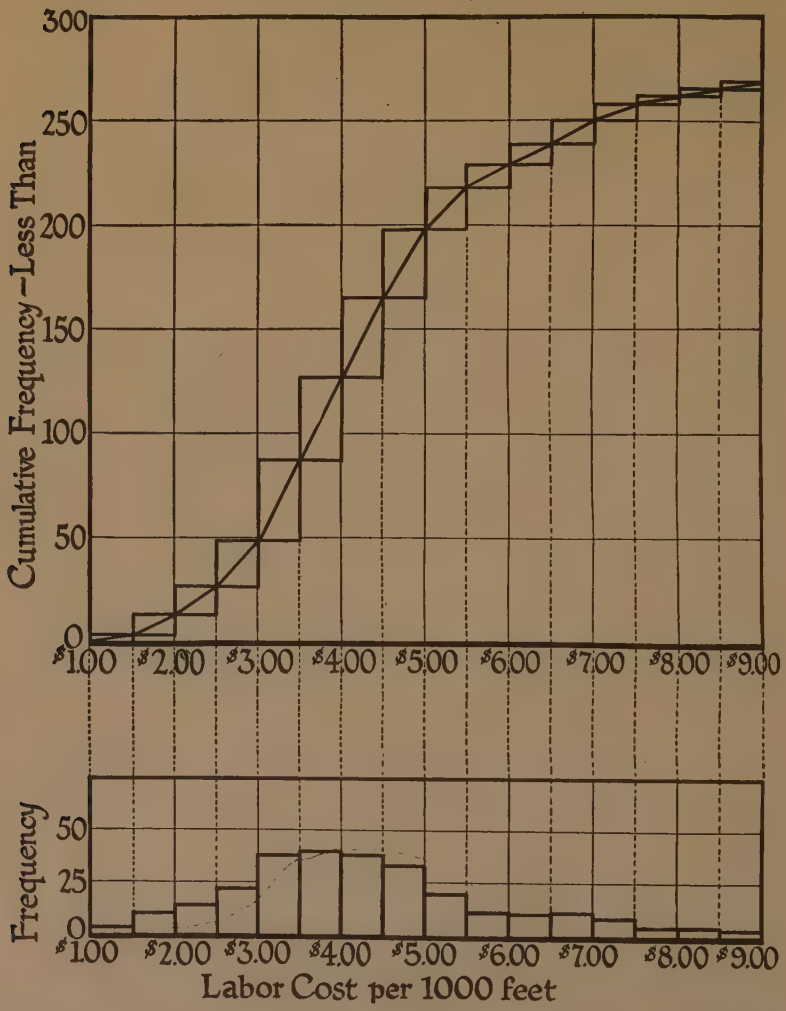


FIG. 35. — Distribution of Sawmills in the United States Classified according to Labor Cost in 1920. Illustrating the Structural Relation between the Ogive and the Frequency Curve.

characteristics of each may be more clearly apparent if the structural relationship between these two curves is understood. This relationship is graphically portrayed in Fig. 35.

This figure is based upon the following frequency table, showing the distribution of sawmills in the United States, classified on the basis of labor cost per 1,000 feet of lumber produced.[1]

TABLE 15

Frequency Distribution of 269 Sawmills in the United States Classified According to Labor Cost in 1921

Labor cost (all employees) per 1,000 feet, board measure	*Number of establishments (frequency)*
$1.00–$1.49	3
1.50– 1.99	10
2.00– 2.49	14
2.50– 2.99	22
3.00– 3.49	38
3.50– 3.99	40
4.00– 4.49	38
4.50– 4.99	33
5.00– 5.49	20
5.50– 5.99	11
6.00– 6.49	10
6.50– 6.99	11
7.00– 7.49	8
7.50– 7.99	4
8.00– 8.49	4
8.50– 8.99	3
	269

The upper part of Fig. 35 indicates the method by which the ogive is built up. Just as in the histogram, the area of each rectangle is proportional to the number of cases falling in the given class. Since the operation is a cumulative one, however, the base of each rectangle is the cumulated frequencies of all preceding classes. Thus the y-value (frequency) of the first rectangle is 3, erected from zero as a base, the y-value of the second class is 10, erected from 3 as a base, and so on. The slope of the curve connecting

[1] From "Labor Efficiency and Productiveness in Sawmills," Ethelbert Stewart, *Monthly Labor Review*, January, 1923, 14. Seven scattered cases above $9.00 in value have been omitted from the table and the accompanying graph.

these rectangles is gradual at first when the frequencies are low, then steeper as the frequencies become greater, and finally tapers off as the frequencies decrease near the upper limit of the distribution. This is the cumulative frequency curve, or ogive.

When the various rectangles representing the class-frequencies are dropped to the zero line as a common base, the x-values remaining the same throughout, the histogram or column diagram described in an earlier section is secured. From this the frequency polygon or smoothed frequency curve may be derived.

REFERENCES

Bowley, A. L. *Elements of Statistics* (52–81).

Chaddock, R. E. *Principles and Methods of Statistics* (Chaps. IV, V).

Day, Edmund E. *Standardization of the Construction of Statistical Tables.* Quarterly Publications, American Statistical Association, March, 1920.

Jones, D. C. *A First Course in Statistics* (5–21).

Kelley, Truman L. *Statistical Method* (1–37.)

King, W. I. *Elements of Statistical Method* (83–107).

Pearl, Raymond. *Medical Biometry and Statistics* (74–104).

Rugg, H. O. *Statistical Methods Applied to Education* (57–94).

Secrist, Horace. *Introduction to Statistical Methods* (116–157).

Secrist, Horace. *Readings and Problems in Statistical Methods* (242–271).

Yule, G. Udney. *An Introduction to the Theory of Statistics* (75–105).

CHAPTER IV

DESCRIPTION OF THE FREQUENCY DISTRIBUTION: AVERAGES

The classification of quantitative data and the construction of a frequency distribution constitute an important stage in the task of organization and analysis. By means of classification the underlying structure of the data may be revealed and the essential unity of a mass of material may be brought out. But this is only the first step in statistical analysis. It remains to develop methods of measuring and expressing more concisely the significant characteristics of a body of data. For certain purposes the frequency distribution itself must be summarized and condensed, must be boiled down until its essence has been distilled into three or four significant figures.

If each frequency distribution constituted a novel and unique problem, obeying a law peculiar to itself, the task of studying and describing such distributions would be a difficult one. Fortunately this is not so. Quantitative data in widely different fields, when assembled in frequency distributions, show certain common characteristics, obey certain general laws. Experience in one field, therefore, constitutes a guide to work in others. Uniformity in the behavior of masses of data makes possible the development of a generalized method of organizing, analyzing and comparing measurements drawn from many fields of scientific study.

Comparison of Frequency Distributions

This fact of a common law of arrangement running through the universe of quantitative facts may be brought home most effectively by a comparison of distributions illustrative of various types of data. The characteristics of the

frequency distributions and of the frequency curves which follow should be noted, and the distributions compared.

The curve in Fig. 36 is based upon the following data relating to the heights of 18,780 soldiers in a certain army.[1]

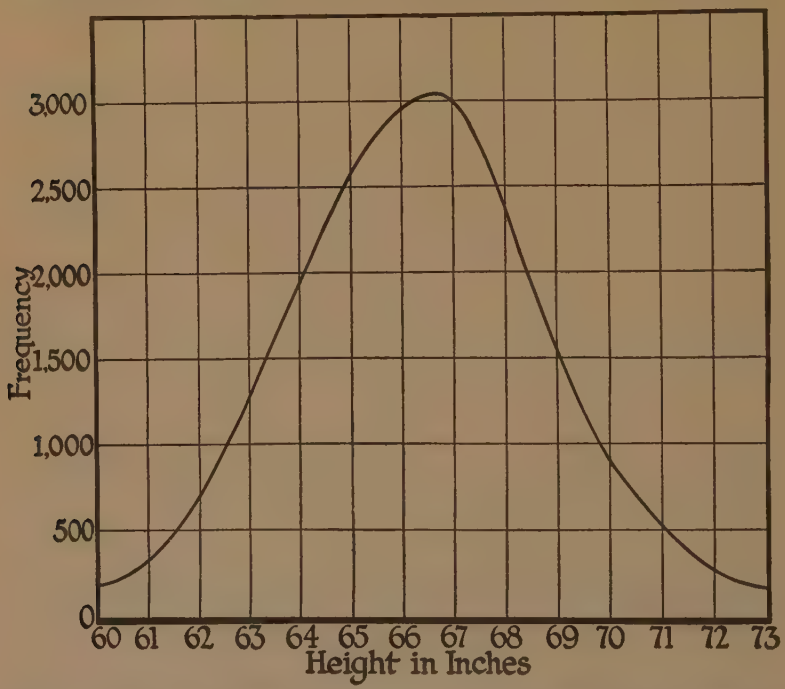


FIG. 36. — Frequency Curve: Distribution of 18,780 Soldiers Classified according to Height

TABLE 16

Distribution of Soldiers Classified According to Height

Height in Inches	*Number of Soldiers*	*Height in Inches*	*Number of Soldiers*
60 +	197	67 +	3,017
61 +	317	68 +	2,287
62 +	692	69 +	1,599
63 +	1,289	70 +	878
64 +	1,961	71 +	520
65 +	2,613	72 +	262
66 +	2,974	73 +	174
		Total	18,780

[1] From G. C. Whipple, *Vital Statistics*, 377.

Figure 37 depicts a frequency curve based upon the errors of observation in 470 astronomical measurements made by Bradley.[1] The data as presented are in the following form:

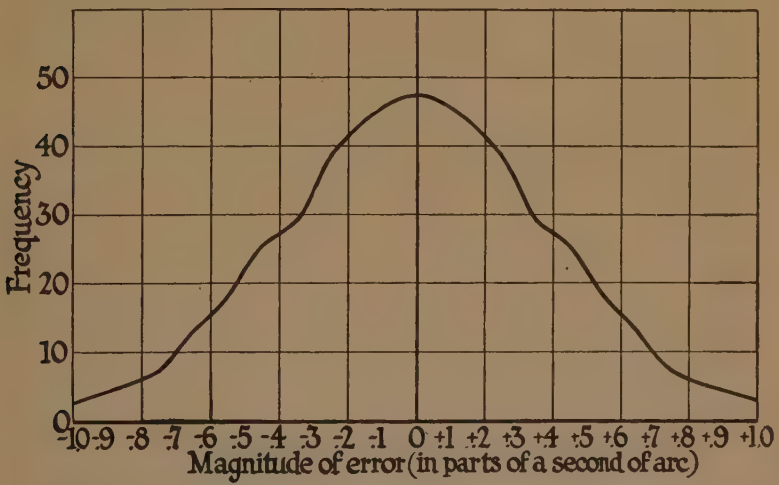


FIG. 37. — Frequency Curve: Distribution of Errors of Observation (Astronomical)

TABLE 17

Distribution of Errors of Observation in Certain Astronomical Measurements

Magnitude of Errors in Parts of a Second of Arc, Between	*Number of Errors of each Magnitude*
0.0 and 0.1	94
0.1 " 0.2	88
0.2 " 0.3	78
0.3 " 0.4	58
0.4 " 0.5	51
0.5 " 0.6	36
0.6 " 0.7	26
0.7 " 0.8	14
0.8 " 0.9	10
0.9 " 1.0	7
above 1.0	8
Total	470

[1] As given by Mellor, *Higher Mathematics for Students of Chemistry and Physics*, 514.

These data, when plotted, will give but one-half of a frequency curve, as no attempt has been made to separate the errors below the true value from those above the true value. The curve has been completed in the figure on the assumption, which is quite valid in this case, that the number of errors of each magnitude below the true value equaled the number of errors of that magnitude above the true value.

If a piece of artillery be accurately adjusted on a given target (a point) and 100 shots be fired, it will be found that the points of impact of the hundred shots will be dispersed about the target. No matter how accurate the piece or the adjustment only a small percentage of the shots will fall upon the exact point at which they were directed. The points of impact will be scattered about the target in a quite regular fashion, however. If a rectangle be so drawn as to include all the points of impact, and this rectangle (or *zone of dispersion*) be divided into eight equal parts, the distribution of shots within these sections will be as indicated in the following diagram. (In any given case there are likely to be slight departures from this order, but in the long run this distribution will prevail.)

2	7	16	25	25	16	7	2

FIG. 38. — Zone of Dispersion, Artillery Firing, Showing the Theoretical Percentage Distribution of Shots

This general rule holds for all classes of guns. The more accurate the gun the smaller will be the zone of dispersion, but the distribution within this zone is theoretically the same in all cases. Rules of fire used in artillery adjustment are based upon this fact.

The results of actual firing may be contrasted with this

theoretical distribution. The accompanying table presents a record of one thousand shots fired from a battery gun at the middle of a stationary target two hundred yards distant.[1] The target was divided by horizontal lines into eleven equal divisions.

TABLE 18

Distribution of One Thousand Shots from a Single Gun

Division	*Number of shots recorded*
1 (top)	1
2	4
3	10
4	89
5	190
6	212
7	204
8	193
9	79
10	16
11 (bottom)	2
Total	1,000

These results are presented graphically in Fig. 39.

The zone of dispersion being divided into eleven divisions instead of the eight referred to in describing the theoretical distribution, a direct comparison cannot be made. We have here, however, the same general type of distribution found in the other examples given. A slight tendency toward concentration in the lower half of the target is undoubtedly due to failure to correct completely for gravitation in the laying of the gun.

When coins are tossed the distribution of heads and tails is assumed to be determined by pure chance. In a single experiment ten coins were tossed 100 times. The following table shows the frequencies with which given

[1] This experiment is recorded in the Report of the Chief of Ordnance, 1878, Appendix S. The results are given in *The Method of Least Squares*, Mansfield Merriman, N. Y., Wiley, 1897, 14.

numbers of heads appeared. (The greatest number of heads possible in a given throw under such conditions is, of course, 10; it is also possible that no heads should appear.)

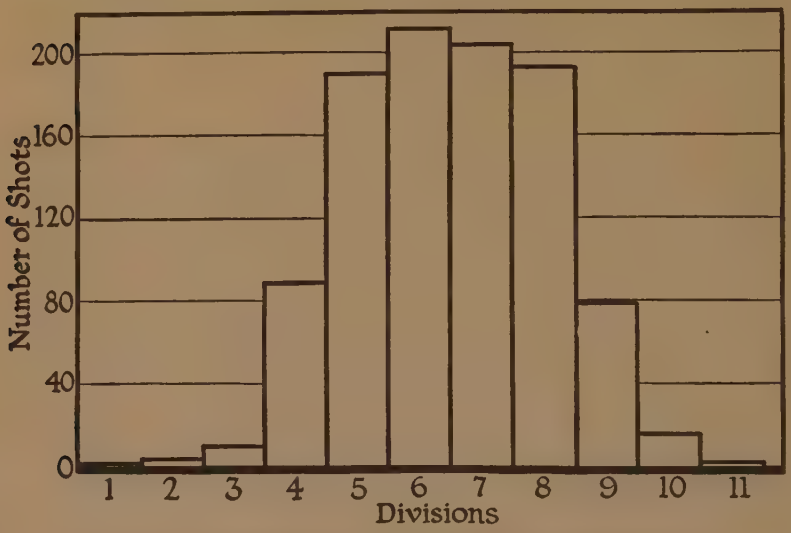


FIG. 39. — Column Diagram: Distribution of 1,000 Shots from a Single Gun

TABLE 19

Distribution of Results in Coin Tossing Experiment
(Ten coins tossed 100 times)

Number of Heads	*Frequency of Occurrence*
10	0
9	1
8	4
7	7
6	23
5	30
4	20
3	9
2	5
1	1
0	0
	100

Figure 40 depicts the above frequency distribution

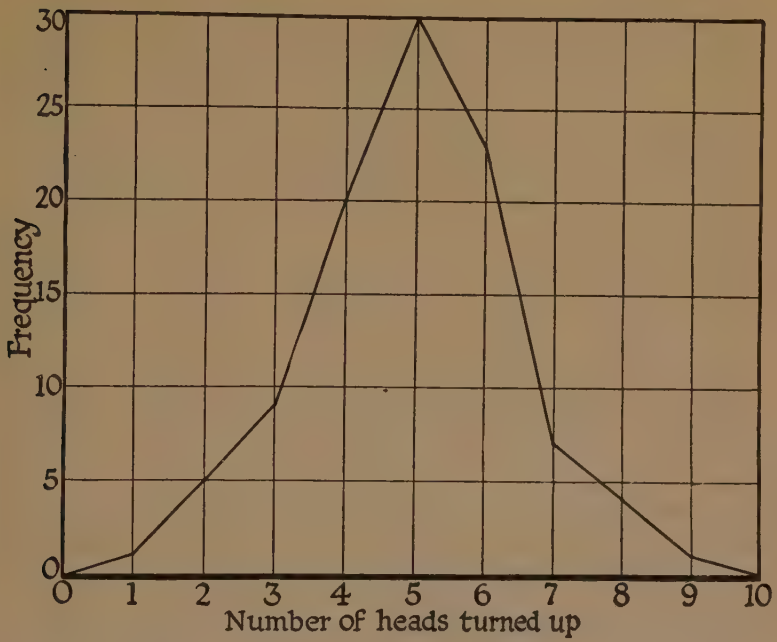


FIG. 40. — Frequency Polygon: Distribution of Heads in a Coin-Tossing Experiment

DISTRIBUTION OF ECONOMIC DATA

We find in these four widely different fields something approaching a uniform law of arrangement of quantitative data. The examples which have been given, however, do not represent the world of economic facts. Do economic data show the same general characteristics? If reference be made to the examples given in Chapter III, comparisons with the four preceding illustrations may be made. The frequency distributions referred to are those relating to weekly earnings of employees, the length of life of telephone poles, the distribution of labor cost in sawmills and the distribution of incomes below $4,000 in the United

States. (The curve of the latter distribution, it should be noted, would show a long tail extending far to the right if the incomes above $4,000 were included.) Several additional examples of economic data may be given.

Figure 41 illustrates the order in which price variations are distributed. It is based upon a study made by

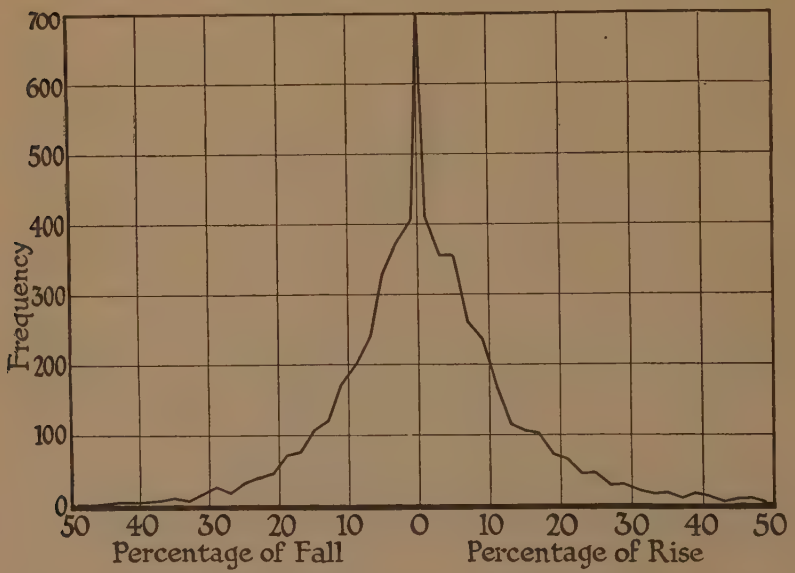


FIG. 41. — Frequency Polygon: Distribution of 5,540 Cases of Change in the Wholesale Prices of Commodities from One Year to the Next (After Mitchell)

W. C. Mitchell of 5,578 individual cases of change in the wholesale prices of commodities from one year to the next.[1] Thus, for example, the average price of middling upland cotton in New York in 1912 was $0.115 per pound. In 1913 the average price was $0.128 per pound, an increase

[1] From *Bulletin* 284, U. S. Bureau of Labor Statistics, Part I, "The Making and Using of Index Numbers," 18. The figure shows the price changes only within the range of a 51% fall and a 51% rise. One case of a price fall of 55% is not shown, and 37 cases of price increases ranging from 52% to 104% have not been included.

of 11.3%. This would constitute one entry in the table of rising prices, falling in the class "10–11.9%." The entire table consists of 5,578 such entries. These data are presented in Fig. 41 in the form of a frequency polygon, no attempt being made to smooth the curve.

Table 20 shows the distribution of London-New York exchange rates (sterling exchange) during the period 1882–1913, inclusive. The time factor has been ignored and

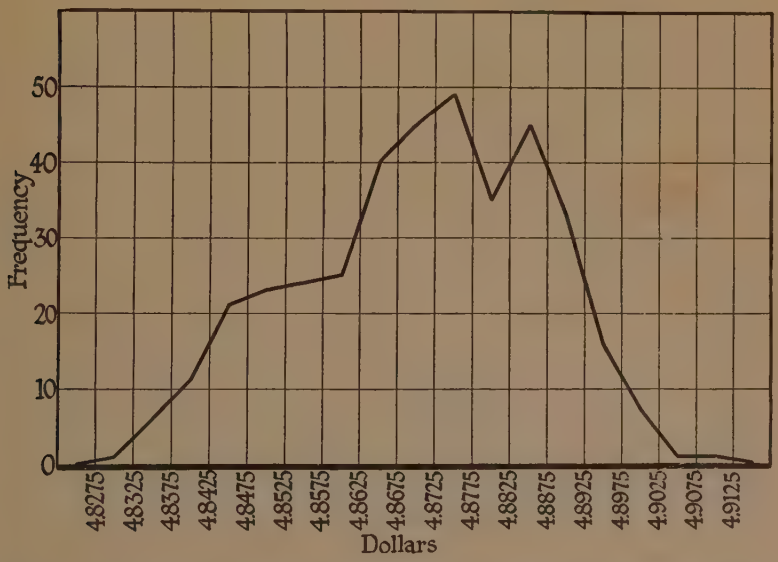


FIG. 42. — Frequency Polygon: Distribution of London-New York Exchange Rates (as recorded over a period of 384 months)

monthly rates have been classified according to the frequency of their occurrence over this thirty-two year period.[1] The data are presented graphically in Fig. 42.

[1] "The figures are ..... the averages of those quoted at the beginning of each month in the *Economist;* on and after July, 1886, the exchange is the 'telegraphic transfer', before that date, 'short at interest.'" The data are taken from *An Academic Study of Some Money Market and Other Statistics,* by E. G. Peake. London, P. S. King, 1923. Appendix I.

TABLE 20

Distribution of London-New York Exchange Rates as Recorded by Months During the Period 1882–1913

Class-interval	*Frequency (Number of months given rate prevailed)*
$4.8275–$4.8324	1
4.8325– 4.8374	6
4.8375– 4.8424	11
4.8425– 4.8474	21
4.8475– 4.8524	23
4.8525– 4.8574	24
4.8575– 4.8624	25
4.8625– 4.8674	40
4.8675– 4.8724	45
4.8725– 4.8774	49
4.8775– 4.8824	35
4.8825– 4.8874	45
4.8875– 4.8924	33
4.8925– 4.8974	16
4.8975– 4.9024	8
4.9025– 4.9074	1
4.9075– 4.9124	1
	384

The frequency curves and histograms based upon economic data, it will be noted, do not all show the symmetry and regularity which seem to characterize the curves representing physical data. Some are skewed to the right or left, there are breaks in the regularity of the increase or decrease of frequencies in certain of the examples. But in spite of these differences there is obviously a family resemblance between the measurements drawn from the fields of economics, astronomy, anthropometry, ballistics, and pure chance. Certain of the common characteristics may be noted.

General Characteristics of Frequency Distributions

There is, in the first place, *variation* in the values of the measurements secured. Human heights vary, astronomical measurements of the same quantity differ, projectiles fired

under conditions as nearly constant as it is humanly possible to make them fail to land at the same spot, incomes vary as between individuals, and exchange rates move from week to week and month to month. The various observations or values secured in a given case are distributed along a scale, between two extreme values.

The distribution of these values along the scale (the x-axis) is such that, moving from one extreme value towards the other, the cases found at successive points along the scale (the successive class frequencies) increase with more or less regularity up to a maximum, and then decrease in much the same way. In spite of variation, therefore, we find a *central tendency*, a massing of cases at certain points on the scale of values. This is the second notable characteristic which all the frequency distributions appear to possess in common.

If we measure, for each of the successive classes, the amount of deviation along the scale from the point of greatest concentration it will be noted that small deviations are much more frequent than large ones, that extreme deviations are rare, and that deviations on both sides of the point of concentration reach perfect (or almost perfect) equality in the examples taken from the physical sciences and from the field of pure chance, and approximate equality in the economic distributions. (Exceptions to this rule of approximate equality on the two sides of the point of greatest concentration are not infrequent, the example of income distribution being a rather striking case in point.)

Figure 43 depicts a curve which is termed the "probability curve," or the "normal curve of error." Its characteristics will be discussed in greater detail in a later section. At this point it is presented merely as a basic type which some of the above examples approach closely, and from which others of the examples represent more or less pronounced deviations. Departures from this type, let it be emphasized, are numerous and significant, but as a basic

form this normal curve of error is extremely important in statistical work. Even the most important variations from this type resemble it with sufficient closeness to justify the use of a generalized method of describing frequency distributions. Distributions of quantitative data vary, and their

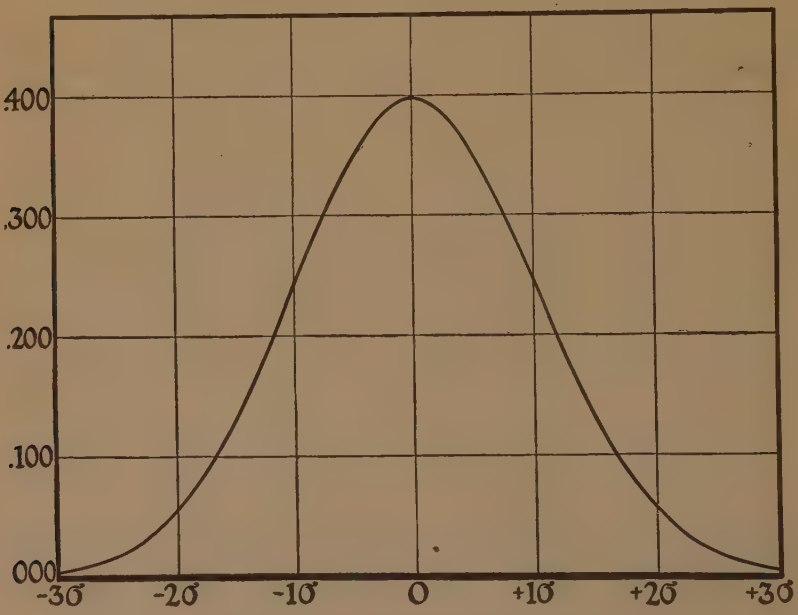


Fig. 43. — The Normal Curve of Error

variations from each other and from certain standard types are of the greatest significance, but in spite of their variations a family resemblance runs through them all. Each new frequency distribution is not an isolated phenomenon, but a member of a large family, and as such the problem of describing and analyzing it may be approached with confidence in methods which have been found applicable in other cases.

Given this more or less common type, how may a given distribution be described and differentiated from others? Certain methods will have been suggested by the preceding discussion.

Methods of Describing a Frequency Distribution

The values of all the observations, it has been noted, are spread along a scale. The frequency distribution may be described by the selection of a single value on that scale which is thoroughly representative of the distribution as a whole. Since the frequencies vary, an obvious choice is the selection of that value which occurs the greatest number of times, or, in other words, that point on the scale at which the concentration is greatest. This value constitutes a *measure of the central tendency* of the distribution. Thus, one might find the income class in which the greatest number of people fall, and let the mid-point of that class (which is $950 in the distribution presented in Table 11) serve as the representative of the distribution. This most common value, it should be noted, is only one of several possible measures of the central tendency of a given distribution. All such measures are termed *averages*.

A single representative value such as this has many uses but, by itself, it obviously leaves out many facts concerning the distribution. Of great importance is the character of the distribution about the average. Are the values of all tabulated cases closely concentrated, or is there pronounced dispersion over a wide range? The representative character of any average depends upon how closely the other values cling to it, upon the degree of concentration about the central tendency. The average, therefore, must be supplemented by a *measure of variation*, a measure of the "scatter" about the central value.

An adequate description should include also an account of the degree of symmetry of the distribution. It is highly important to know whether there is an equal distribution of cases on each side of the point of greatest concentration, or whether the frequency curve is skewed to one side, as in the case of income distribution illustrated above.

If the curve is not symmetrical the degree of asymmetry should be determined, and for this purpose *measures of skewness* have been developed.

It is, finally, possible to measure the degree of peakedness of frequency curves, by comparing them with the normal curve of error as a standard. It is obvious that the frequency polygon representing price changes (Fig. 41) would, if smoothed, constitute a curve much more peaked than the normal curve, and this fact of pronounced concentration at the central value is highly significant. This characteristic of frequency curves is called *kurtosis*, and the measurement of kurtosis constitutes the final step in the description of the frequency distribution.

When these various measures have been secured the task of statistical analysis will be well under way. The chaotic assortment of data with which we started will have been reduced to workable form in the shape of a frequency table, and the essential facts which the table reveals will have been distilled into three or four significant measures. This process not only reveals the characteristics of the given distribution, but also facilitates comparison with similar distributions. For example, it is impossible to compare some tens of millions of unorganized personal income figures for the United States with similar data for Great Britain. But if we secure a value for the average or most representative income for each country, together with a description of the distribution of personal incomes about that central value, a legitimate basis for comparative study is obtained. In manipulating and analyzing masses of material, whatever the purpose of study may be, full use should be made of the power to condense, simplify and compare which is given by the measures employed in describing the frequency distribution.

The succeeding section is devoted to a discussion of one phase of this descriptive process, that concerned with the measurement of central tendencies. After the development

of this subject of averages, problems relating to measures of variation and of skewness will be dealt with.

Averages

We have seen that the representation of a frequency distribution by an average, a single typical figure, is justified because of the tendency of large masses of figures to cluster about a central value, from which the values of all observed cases depart with more or less regularity and smoothness. It is solely because of the concentration of cases about a central point on the scale that such representative figures have significance. The average represents the distribution as a whole only because it is a typical value. If the individual items entering into a distribution vary widely in value and show no tendency toward concentration, no single value can represent them. Thus the arithmetic mean of the three numbers 3; 125; 1000 is 376, but 376 in no way represents the three values on which it is based. This fundamental requirement, that there be a tendency toward concentration about a central value, must be met if an average is to be at all representative.

If the general character of a frequency distribution be recalled the logic of one sort of average will be clear at once. It was suggested above that that point on the x-scale at which the concentration is greatest, that value which occurs the greatest number of times, might be taken as typical of the entire distribution. This value is termed the *mode*, and the group in which it falls is called the *modal group*. If a frequency curve be drawn to represent a given distribution, the mode will be the *x-value corresponding to the maximum ordinate*.[1] The maximum ordinate itself measures the frequency of the modal group. Students frequently confuse these two values in determining the mode.

[1] Strictly speaking, the mode is the x-value corresponding to the maximum ordinate of the ideal frequency curve which has been fitted to the given distribution.

It is not the distance along the y-scale but the distance along the x-scale which measures the value of the mode. The ordinates merely measure the number of cases falling in the several classes, not the values of the cases falling in those classes.

As typical of a given distribution we might also select that point on the scale of x-values on each side of which one half the total number of cases fall. This value, which is called the *median*, is that which exceeds the values of one half the cases included, and is in turn exceeded by the values of one half the cases. Thus it has been estimated that in 1918 the median value of personal incomes in the United States was $1,140; one half of the 37 million recipients of personal incomes received less than this sum, while one half received more. When a distribution is represented by a frequency curve, the area under the curve is divided into two equal parts by an ordinate erected at that point on the x-axis corresponding to the median value. This follows, of course, from the definition of the median, and from the fact that the area under a frequency curve represents the total number of cases included in the distribution.

The *arithmetic mean* is a third type of average which may be used to represent a distribution. This is a *calculated* average, affected by the value of every item in the distribution. Herein, obviously, it differs from the mode and the median, which depend primarily upon the relative position of the items in the frequency table, and are not affected by the values of all individual items. The arithmetic mean is the center of gravity of a distribution; it would be the x-value of the point of balance of a frequency curve, if the curve could be blocked out and manipulated in solid form.

The *geometric mean* and the *harmonic mean* are two other averages the characteristics of which will be discussed at a later point.

The computation or location of these various averages

may involve somewhat lengthy processes if the number of cases included be great. If appropriate methods be employed, however, the labor of computation may be materially cut down. The use of the following symbols will simplify the explanation of these methods:

M: Arithmetic mean.

Mo: Mode.

Md: Median.

m: The value of an individual observation; in a frequency distribution, the value of the midpoint of a class.

f: The number of cases in a given class in a frequency distribution.

N: The total number of cases in a given series or frequency distribution.

Σ (Sigma): The symbol for the process of summation, meaning "the sum of."

The Computation of the Arithmetic Mean

Using the above notation, the formula for the arithmetic mean is

$$M = \frac{\Sigma m}{N}.$$

Thus the mean of the measures 2, 5, 6, 7, is equal to the sum of these measures divided by 4, which is $\frac{20}{4}$ or 5. The computation of the arithmetic mean when each measure is reported at its true value is thus a simple process of summation and division. The weekly earnings of 210 factory employees were listed in an earlier section. If these figures be added, and the total divided by 210, the mean weekly wage is found to be \$26.982. In this case the task of adding 210 items is somewhat tedious; it is a task which would become almost impossible if one were dealing with the 37 million personal income figures, for example. For

practical reasons, therefore, it is usually necessary to compute the required averages from the frequency distribution rather than from the original ungrouped data. To exemplify this process certain data relating to the yield of wheat in North Dakota may be utilized. The average yield of wheat per acre in each of the 53 counties of North Dakota for the period 1911–1921 has been recorded by the U.S. Bureau of Agricultural Economics. The figures are summarized in the following table:

TABLE 21

Calculation of the Arithmetic Mean of Wheat Yield in the Counties of North Dakota, 1911–1921 [1]

Class-interval (in bu. per acre)	*Mid-point* m	*Frequency* f	fm
0– 1.9	1	3	3
2– 3.9	3	26	78
4– 5.9	5	78	390
6– 7.9	7	107	749
8– 9.9	9	113	1,017
10–11.9	11	65	715
12–13.9	13	40	520
14–15.9	15	22	330
16–17.9	17	45	765
18–19.9	19	41	779
20–21.9	21	21	441
22–23.9	23	8	184
		569	5,971

$$M = \frac{\Sigma(fm)}{N} = \frac{5{,}971}{569} = 10.49 \text{ bu.}$$

The importance of certain of the precautions mentioned in the section on classification will be clear from this example. It was there stated that the class-interval should be so selected that the error in assuming an even distribution of cases throughout each class-interval would not be great. If the cases in each class are evenly distributed we

[1] Average yields, by counties, for this period are given in *Bulletin* 165, Agricultural Experiment Station, North Dakota Agricultural College, 1922, *Cost of Production and Farm Organization on 126 Farms in North Dakota, 1921,* 120–21.

may take the mid-point of that class as representative of all the cases included, but if there is not an even distribution the mid-point of the class ceases to be representative. The process of calculating the value of the mean from the frequency distribution rests upon the assumption of an even distribution of cases within each class. The mid-point of each class is multiplied by the number of cases in that class, it being assumed that the product is approximately equal to the sum of all the individual items in the class. The formula for the mean thus becomes $M = \frac{\Sigma(fm)}{N}$. Table 21 illustrates the procedure in detail.

The value secured in this way is sometimes called a weighted arithmetic mean. What we do, in effect, is to secure the arithmetic mean of the 12 figures in the column headed *m*. We do not take a simple average of these figures, however, but *weight* each one in proportion to the number of cases falling in the class-interval of which it is the mid-point. It is precisely the same procedure as that which we should follow in calculating the mean of five men's incomes, two of whom, let us say, have incomes of $2,000 and three of whom have incomes of $3,000. Clearly it would not do to add the figures $2,000 and $3,000, dividing the sum by two. The figure $2,000 is given a weight of 2, the figure $3,000 is given a weight of 3, and the resultant sum, $13,000, is divided by 5. Though the procedure in working from the frequency distribution is thus a form of weighting, the term "weighted average" is coming to have a more restricted meaning, to be explained at a later point, and should not in general be applied to an average computed from a frequency distribution.

Short Method of Computing the Arithmetic Mean

The calculation of the arithmetic mean from the frequency table is much easier, in general, than from the un-

grouped data, but when the number of cases included is large even the computation from the frequency table by the method illustrated above may be laborious. The procedure may be greatly simplified.

From the method of computing the arithmetic mean it follows that the algebraic sum of the deviations of a series of individual magnitudes from their mean is zero. This may be readily demonstrated. We may represent the series of magnitudes by m_1, m_2, m_3, . . . m_n, their arithmetic mean by M, and the deviations of the various magnitudes from the mean by d_1, d_2, d_3, . . . d_n.

Then

$$M = \frac{m_1 + m_2 + m_3 + \ldots + m_n}{N} \quad (1)$$

and

$$NM = m_1 + m_2 + m_3 + \ldots + m_n \quad (2)$$

The number of terms, of course, is equal to N. Therefore

$$(m_1 - M) + (m_2 - M) + (m_3 - M) + \ldots + (m_n - M) = 0 \quad (3)$$

But

$m_1 - M = d_1$, $m_2 - M = d_2$, etc., and equation (3) may be written

$$\Sigma d = 0.$$

Knowing this to be true we may measure the deviations of a series of magnitudes from any arbitrary quantity, secure the algebraic sum of the deviations, and from this value ascertain the difference between the arbitrary quantity and the true mean. For this difference will be the mean of the deviations from the arbitrary origin. If we let M' represent the arbitrary origin, or assumed mean, while $c = M - M'$, and d_1', d_2', d_3' . . . d_n', represent the deviations of the various magnitudes from M' (i.e. $d_1' = m_1 - M'$, $d_2' = m_2 - M'$, etc.), then

$$d_1' = d_1 + c,\ d_2' = d_2 + c,\ d_3' = d_3 + c,\ \ldots\ d_n' = d_n + c$$

and

$$\Sigma d' = \Sigma d + Nc$$

But

$$\Sigma d = 0$$
$$\therefore \Sigma d' = Nc$$

and

$$c = \frac{\Sigma d'}{N}$$

From the known values of M' and c the value of the true mean may be obtained, for $M = M' + c$. The procedure is illustrated in the following simple example:

TABLE 22

Computation of the Arithmetic Mean (Short Method)

(Ungrouped data)

m	f	d'
5	1	− 15
15	1	− 5
25	1	+ 5
35	1	+ 15
45	1	+ 25
	5	+ 25

$M' = 20$

$$c = \frac{\Sigma d'}{N} = \frac{+25}{5} = 5$$

$$M = M' + c = 20 + 5 = 25$$

When the deviations are measured from 20 as arbitrary origin there is in each case a constant error, if the deviation from the true mean be taken as standard. This error is equal to the difference between the true and the assumed means. The algebraic sum of the deviations from the assumed mean will equal N times this constant error, since the error is repeated once for every item included. By dividing the sum of these deviations by N the amount of the error may be determined and the value of the mean thus obtained.

The work of computation may be still further abbreviated in the case of quantities arranged in the form of a frequency distribution by measuring the deviations in terms of the class-interval as a unit. Then, in finally applying the necessary correction, the difference between the true and assumed means may be again expressed in terms of the original units. The method may be illustrated in

detail by means of the wheat yield data for which the mean has already been calculated.

TABLE 23

Calculation of the Arithmetic Mean of Wheat Yield in the Counties of North Dakota, 1911–1921

(Short Method)

Class-interval (in bu. per acre)	*Mid-point m*	*Frequency f*	*d′ (in class-interval units)*	*fd′* −	*fd′* +	*Calculations* $M' = 9$
0 – 1.9	1	3	− 4	12	.....	1. Algebraic sum of deviations from M'
2 – 3.9	3	26	− 3	78	.....	+ 778
4 – 5.9	5	78	− 2	156	.....	− 353
6 – 7.9	7	107	− 1	107	.....	+ 425
8 – 9.9	9	113	0	.....	.....	2. Calculation of c (in class-interval units):
10 – 11.9	11	65	1	.....	65	$c = \frac{425}{569} = .7469$
12 – 13.9	13	40	2	.....	80	3. Reduction of c to original units:
14 – 15.9	15	22	3	.....	66	Class-interval = 2
16 – 17.9	17	45	4	.....	180	c (in original units)
18 – 19.9	19	41	5	.....	205	$= .7469 \times 2$
20 – 21.9	21	21	6	.....	126	$= 1.4938$
22 – 23.9	23	8	7	.....	56	4. Determination of M
		569		− 353	+ 778	$M = M' + c$
						$M = 9 + 1.4938$
						$M = 10.4938$ bu.

The steps in this process of calculating the arithmetic mean by the short method may be briefly summarized:

1. Organize the data in the form of a frequency distribution.
2. Adopt as the assumed mean the mid-point of a class near the center of the distribution.
3. Arrange a column showing the deviation (d') from the assumed mean of the items in each class, in terms of class-interval units. This deviation will be zero for the

items in the class containing the assumed mean, -1 for the items in the next lower class, $+1$ for the items in the next higher class, and so on.

4. Multiply the deviation of each class by the frequency of that class, taking account of signs. These products are entered in the column fd'.
5. Get the algebraic sum of the items entered in the column fd'.
6. Divide this sum by the total frequency (N). The quotient is the correction (c) in class-interval units.
7. Multiply the correction (c) by an amount equal to the class-interval. The product is the correction in terms of the original units.
8. Add this correction (algebraically) to the assumed mean (M'); the sum is the true mean (M).

Location of the Median: Ungrouped Data

The median is a value of the x-variable so selected that 50 per cent of the total number of cases, when arranged in order of magnitude, lie below it and 50 per cent above it. For many frequency distributions this is a useful and significant value.

When handling data which are not arranged in the form of a frequency distribution the location of the median is a simple matter. The data having been arranged in order of magnitude, it is necessary only to count from one end until that point on the scale of values is found which divides the number of cases into two equal parts. As a simple example we may assume the following seven figures to represent the annual incomes of seven individuals:

\$750 \$975 \$1128 \$1450 \$1475 \$1825 \$1950

The scale of values extends from \$750 to \$1950, and seven items are arranged along this scale. The value of \$1000 has two items on one side and five items on the other, so obviously does not conform to our definition of

the median. The value of $1450, which corresponds with the income of one of the seven individuals, is the median in this case. Three items lie on each side of this value; or, if we assume the central item to be cut in two, $3\frac{1}{2}$ items lie on each side of this point. This case is illustrated in Fig. 44. This diagram may help to bring out the fact that the median is a point on a scale, so located that it cuts the frequencies in two.

The problem is slightly different when an even number of cases is included. This condition is exemplified in the following table which shows the average earnings of band sawyers, in sawmills in each of 22 States, during the year 1921.

TABLE 24

Average Earnings of Band Sawyers by States, 1921 [1]

State	*Average earnings per one-man hour*
Virginia	$0.586
Pennsylvania	.621
Tennessee	.627
North Carolina	.647
West Virginia	.658
Maine	.686
South Carolina	.721
Wisconsin	.729
Michigan	.730
Alabama	.749
Minnesota	.755
Texas	.761
Georgia	.779
Arkansas	.785
Mississippi	.798
Louisiana	.824
Florida	.825
Idaho	.833
California	.864
Montana	.904
Washington	1.045
Oregon	1.106

[1] From *Monthly Labor Review*, January, 1923, 9.

The median must be in this case a value on each side of which 11 States lie. Therefore any value exceeding $0.755 (average earnings in Minnesota) and less than $0.761 (average earnings in Texas) will satisfy the definition of the median. Under these conditions, where the median is really indeterminate, a value half way between the two

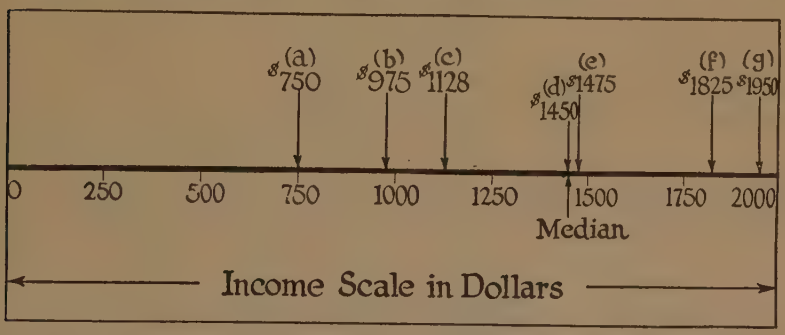


FIG. 44. — Illustrating the Location of the Median with Ungrouped Data (Personal incomes of seven individuals)

limiting values is accepted, by convention. The median of the 22 figures given would thus be $0.758.

In this example the median value does not correspond with the earnings in any one State. This will frequently be so when there is an even number of cases. When the data being handled constitute a discrete series an unreal value, inconsistent with the nature of the data, may be secured for the median. Thus the median value of the discount rates on 60–90 day commercial paper in the New York market during the period 1900–1922 is found to be 4.865 per cent, a thoroughly artificial rate which actually never prevailed in the market.

LOCATION OF THE MEDIAN: GROUPED DATA

The task of locating the median is essentially the same when the data are in the form of a frequency distribution. The fact that the real values of the individual items are not known, because of the grouping by classes, complicates

the problem slightly. The following data, representing the average discount rate on 60–90 day commercial paper in the New York market for 276 consecutive months, may be used in illustrating the method.

TABLE 25

Location of Median, Discount Rates, New York

60–90 da. choice double name commercial paper, 1900–1922

Class-interval (*Discount rate in %*)	*Mid-point* m	*Frequency* f
2.75–3.24	3.00	8
3.25–3.74	3.50	31
3.75–4.24	4.00	52
4.25–4.74	4.50	38
4.75–5.24	5.00	39
5.25–5.74	5.50	42
5.75–6.24	6.00	35
6.25–6.74	6.50	13
6.75–7.24	7.00	5
7.25–7.74	7.50	4
7.75–8.24	8.00	9
		276

$$\frac{N}{2} = \frac{276}{2} = 138$$

$$Md = 4.75 + (9/39 \times .50)$$

$$= 4.75 + .115$$

$$= 4.865$$

In the present case the location of the median involves the determination of that value on each side of which 138 items lie. We may assume that we start at the lower end of the scale and move through the successive classes. When we reach the upper limit of the first class (that including items having values from 2.75 to 3.25) we have left behind us 8 cases, while 268 lie in front of us. When the upper limit of the second class is attained, 39 items have been passed; when the upper limit of the fourth class is reached, 129 items have been passed. The upper limit of the fifth class has below it 168 items. Somewhere between the lower and upper limits of the fifth class lies the desired point, that which has 138 items on each side of it. How far must we move through this class, from 4.75 to 5.25 in order to reach this point?

It will be recalled that, for purposes of calculation, the assumption is made that there is a uniform distribution of

the items lying within any given class. Since before we reach the fifth class 129 cases have been counted, only 9 of the 39 included in this class are needed to complete the desired number, 138. On the assumption of even distribution the required 9 cases will lie within a distance on the scale equal to 9/39 of the class-interval. The class-interval is .50; 9/39 of .50 is equal to .115. As we move up the scale, then, having reached 4.75, we proceed an additional distance equal to .115. At a point on the scale having a value of 4.865 is the dividing line on each side of which lie 138 cases. This is the value of the median.

The process of computation is shown at the right of the frequency table. The following is a summary of the steps involved in the location of the median:

1. Arrange the data in the form of a frequency distribution.
2. Divide the total number of measures by 2; this gives the number which must lie on each side of the point to be located.
3. Begin at the lower end of the scale and add together the frequencies in the successive classes until the lower limit of the class containing the median value is reached.
4. Determine the number of measures from this class which must be added to the frequencies already totaled to give a number equal to $N/2$.
5. Divide the additional number thus required by the total number of cases in the class containing the median. This indicates the fractional part of the class-interval within which the required cases lie.
6. Multiply the class-interval by the fraction thus set up.
7. To the lower limit of the interval containing the median add the result of the multiplication process indicated in (6). This gives the value of the median.

The last three steps constitute merely a simple form of interpolation.

The entire process may be reversed by beginning at the upper end of the scale and counting downwards. In this case the final operation is one of subtraction from the upper limit of the interval containing the median.

If N is an odd number $N/2$ will contain a fractional value. The operation is precisely the same as that outlined above.

Quartiles and Deciles

For many purposes it is desirable to locate on the scale of values, along which the items constituting a frequency distribution are ranged, points dividing the total number of measures in other ways. Similar to the median, which divides the total number of cases into two equal groups, are the quartiles, deciles and percentiles. The quartiles, as the term implies, are points on the scale which divide the entire number of measures into four equal groups, the deciles divide the number into ten equal groups, and the percentiles divide the total number of cases into 100 equal groups. Thus the first quartile is that point on the scale below which one quarter of the total number of cases lie and above which three quarters of the total number of cases lie. The second quartile and the median are identical values. The third decile is that point on the scale below which three tenths of the total number of cases lie and above which seven tenths of the total number of cases lie. In all cases the count begins at the lower end of the scale.

Example. Location of the First Quartile (Q_1), Discount Rates
(See Table 25.)

$$N/4 = 69$$
$$Q_1 = 3.75 + (30/52 \times .50)$$
$$= 4.038$$

Example. Location of Fourth Decile, Discount Rates
(See Table 25.)

$$N/10 = 27.6 \qquad D_4 = 4.25 + 19.4/38 \times .50$$
$$4N/10 = 110.4 \qquad = 4.505$$

A method of locating median, quartiles, deciles and percentiles graphically is explained below.

Location of The Mode

The mode is the value of the x-variable corresponding to the maximum ordinate of a given frequency curve. The

concept of a modal value is a thoroughly easy one to grasp. It is the most common wage, the most common income, the most common height. It is the point where the concentration is greatest, a characteristic which is effectively brought out by Fechner's term for this average, *dichtester wert*, or *thickest value*. It is not so easy, however, to locate the true modal value in a given case. In general statistical work an approximate value only is secured for the mode, but for most practical purposes this value is usually sufficiently accurate.[1]

The method of determining this approximate modal value may be illustrated by reference to the following distribution:

TABLE 26

Frequency Distribution of Five Per Cent Bonds

(The table is based upon quotations on the New York Stock Exchange on Oct. 31, 1923, on railroad and industrial bonds with coupon rate of 5 per cent)

Quoted Price Class-interval	*Mid-point m*	*Frequency f*
Less than 71.5		30
71.5– 73.49	72.5	2
73.5– 75.49	74.5	5
75.5– 77.49	76.5	5
77.5– 79.49	78.5	5
79.5– 81.49	80.5	12
81.5– 83.49	82.5	13
83.5– 85.49	84.5	15
85.5– 87.49	86.5	11
87.5– 89.49	88.5	17
89.5– 91.49	90.5	17
91.5– 93.49	92.5	33
93.5– 95.49	94.5	45
95.5– 97.49	96.5	51
97.5– 99.49	98.5	59
99.5–101.49	100.5	23
101.5–103.49	102.5	9
103.5–105.49	104.5	1
		353

[1] A method of locating the mode more accurately is explained in the note appended to Chapter XV.

There is a wide dispersion of the 30 cases falling below 71.5, and the existence of this "open-end" class makes it impossible to compute the value of the mean, as the table stands. The mode is, therefore, an appropriate average to employ in the present instance.

The class having limits of 97.5 – 99.49 contains the greatest number of cases. This appears to be the modal group, and the mid-point of this class, 98.5, may be tentatively accepted as the value of the approximate mode. But with different classifications quite different values might be secured for the mode. When the original bond quotations are tabulated with varying class-intervals, the following results are secured. (Only the frequencies of the central classes are shown. It is not necessary, for this purpose, to present each of the tables as a whole.)

(*a*)		(*b*)		(*c*)	
Class-interval = 1		*Class-interval* = 2		*Class-interval* = 3	
Class-interval	*f*	*Class-interval*	*f*	*Class-interval*	*f*
93.5– 94.49	17	94.5– 96.49	54	91.5– 94.49	50
94.5– 95.49	28	96.5– 98.49	50	94.5– 97.49	79
95.5– 96.49	26	98.5–100.49	52	97.5–100.49	77
96.5– 97.49	25				
97.5– 98.49	25				
98.5– 99.49	34				
99.5–100.49	18				
100.5–101.49	5				

With a class-interval of 1 a value of 99 is secured for the mode; with a class-interval of 2, but with different class limits than in Table 26, a value of 95.5 is obtained. With a class-interval of 3 a modal value of 96 is secured. Further changes in classification would give still other values. The mode thus appears to be a curiously intangible and shifting average. Its value, for the same data, seems to vary with changes in the size of the class-interval and in the location of the class-limits.

These difficulties arise primarily from limitations to the size of the sample being studied. The true mode, that

value which would occur the greatest number of times in an infinitely large sample, could be located exactly if we could increase indefinitely the number of cases included. For, given sufficient cases, the approximate mode approaches the true mode as the class-interval decreases. Grouping in large classes obscures details, and as these classes are reduced in size more of the details are seen and a truer picture of the actual distribution is secured. But since most practical work is necessarily based upon relatively small samples, the increase in the number of classes reveals gaps and irregularities, and causes such a loss of symmetry and order that doubt arises as to where the point of greatest concentration really lies. The different tabulations of bond prices furnish an excellent example of this.

By mathematical methods it is possible to obtain a value for the true mode without securing an infinite number of cases. The smoothing process has been briefly explained. One sort of smoothing is that which involves the fitting of an appropriate type of ideal frequency curve to the data of a given frequency distribution. This gives, theoretically, the distribution which would be secured by the process first indicated, that of decreasing indefinitely the size of the class-interval and increasing indefinitely the number of cases. The value of the x-variable corresponding to the maximum ordinate of this ideal fitted curve is the true mode.

For most practical purposes approximate values of the mode are adequate, and these may be secured by much simpler methods. A first and rough approximation may be obtained by taking the mid-value of the class of greatest frequency, a method suggested above. If the general rules for classification which were outlined in an earlier section have been followed, this procedure will not generally involve a gross error.

It is possible, given a fairly regular distribution, to secure, by a process of interpolation within the modal

group, a closer approximation than is obtained by accepting the mid-value of this group as the mode. Referring again to the tabulation of bond prices in Table 26 it will be noted that the distribution on the two sides of the modal class is not symmetrical. The modal class is that with a mid-value of 98.5. The class next below, with a mid-value of 96.5, contains 51 cases, while that next above, with a mid-value of 100.5, contains but 23 cases. The disproportion is continued in the succeeding classes below and above, more cases being bulked below the modal class than above. For other purposes we have assumed an even distribution of cases between the upper and lower limits of each class, but it is probable that this is not true of the modal class in the present case. Judging from the distribution outside this class, it is likely that the concentration is greatest in the lower half of the class-interval, that is, between 97.5 and 98.5. The mode, therefore, probably lies below the mid-value, 98.5, rather than precisely at that point. We may attempt to locate it within the group by weighting, assuming a pull toward the lower end of the scale equal to 51 (the number in the class next below) and a pull toward the upper end of the scale equal to 23 (the number in the class next above). This may be expressed by a formula, employing the following symbols:

l = lower limit of modal class.
f_1 = frequency of class next below modal class in value.
f_2 = frequency of class next above modal class in value.
i = class-interval.

The interpolation formula is

$$Mo = l + \frac{f_2}{f_2 + f_1} \times i$$

Applying this formula to the bond price data presented in Table 26, we have

$$Mo = 97.5 + \frac{23}{74} \times 2 = 97.5 + .62 = 98.12$$

A closer approximation may sometimes be secured by basing the weights (represented by f_2 and f_1) upon the total frequencies of the two or three classes next above the modal class and the same number below. If three classes on each side are included in the present case, a value of 97.91 is secured for the mode of bond prices.

In some cases the problem of locating the mode is complicated by the existence of several points of concentration, rather than the single point which has been assumed in the preceding explanation. Thus in Table 9, representing the distribution of wages, with a class-interval of 25 cents, there are two definite modal points. A distribution of this type is called bi-modal; when plotted, a frequency curve having two humps is obtained. If the data are homogeneous such a distribution is the result of paucity of data and of the method of classification employed. It may be due to the use of a class-interval too small, with respect to the number of cases included in the sample. An approximate mode may be determined in such cases by shifting the class-limits and increasing the class-interval, carrying on this process until one modal group is definitely established. This reverses the process by which the true mode may be located when the number of cases is infinitely large. Under such conditions the class-interval might be reduced until it was infinitely small. But with a limited number of cases the location of the point where the concentration is greatest necessitates increasing the size of the class-interval, in order to get away from the irregularities due to the smallness of the sample.

If the distribution remains bi-modal in spite of changes in the class-intervals and class-limits, it is probable that the data are not homogeneous, that two different distributions have by mistake been combined. Such cases are not uncommon in biometrical work. The existence of two distinct animal species where only one was suspected has been revealed in this way. The whole significance of a

frequency distribution will be lost if the data are not homogeneous, a fact which is as true of work in the field of economic statistics as in any other.

Determination of the Modal Value from Mean and Median

Another method of securing an approximate value for the mode, a method based upon the relationship between the values of the mean, median and mode, may be employed in certain cases. In a perfectly symmetrical distribution mean, median and mode coincide. As the distribution departs from symmetry these three points on the scale are pulled apart. If the degree of asymmetry is only moderate the three points have a fairly constant relation. The mode and mean lie farthest apart, with the median one third of the distance from the mean towards the mode. If the asymmetry is marked, no such relationship may prevail. Having the values of any two of the averages in a moderately asymmetrical frequency distribution, therefore, the other may be approximated. In fact, however, the method should only be employed in determining the value of the mode, as the other two values may be computed more accurately by other methods. The value of the mode itself should only be determined in this way when more exact methods are not applicable or are not called for.

The following formula is based upon this relationship:

$$Mo = \text{Mean} - 3\,(\text{Mean} - Md)$$

Applying this formula to the telephone pole data shown in Table 12, the following result is secured:

$$Mo = 9.33 - 3(9.33 - 9.015) = 8.385$$

This value is slightly below the mid-value of the modal class, 8.5, and is also less than the value of 8.49 which is secured by weighting within the modal group (using four classes on each side).

It must be emphasized that there is a fictitious accuracy to all these values for the mode. All the methods of locating the mode which have been discussed are merely approximative, a fact which must not be forgotten in interpreting and utilizing the results.

Graphic Location of Mode, Median, Quartiles and Deciles

A better understanding of the frequency curve and of the cumulative frequency curve may be secured through a brief discussion of certain methods of locating graphically some of the statistical measures which have been described.

The value of the mode may be readily determined from a frequency curve of the usual type, for, by definition, the mode is the reading on the horizontal scale corresponding to the maximum ordinate of such a curve. If this reading be taken from the frequency polygon a rough value will be obtained, the mid-value of the class of greatest frequency. A closer approximation to the true value of the mode will be secured from a curve which has been smoothed, either by inspection or by mathematical methods. Figure 45, showing a curve (smoothed by inspection) based upon the wage data presented in Table 8, indicates how the mode may be located graphically. The horizontal reading corresponding to the maximum ordinate of this curve is $27.50, an approximate value of the mode which may be compared with the values of $27.69 secured by the weighting process and of $27.3470 secured from the values of the mean and median.

The locations of the median and mean have been indicated on this curve. It has been pointed out that in moderately asymmetrical (or skewed) distributions there tends to be a constant relationship between the three averages which have been described, the median lying between the mean and the mode, and approximately one third of the

distance from the former towards the latter. In the present case this relationship holds fairly well when the value of the mode is approximated from the smoothed curve. The irregularities in the original data render the process of smoothing by inspection rather arbitrary, however.

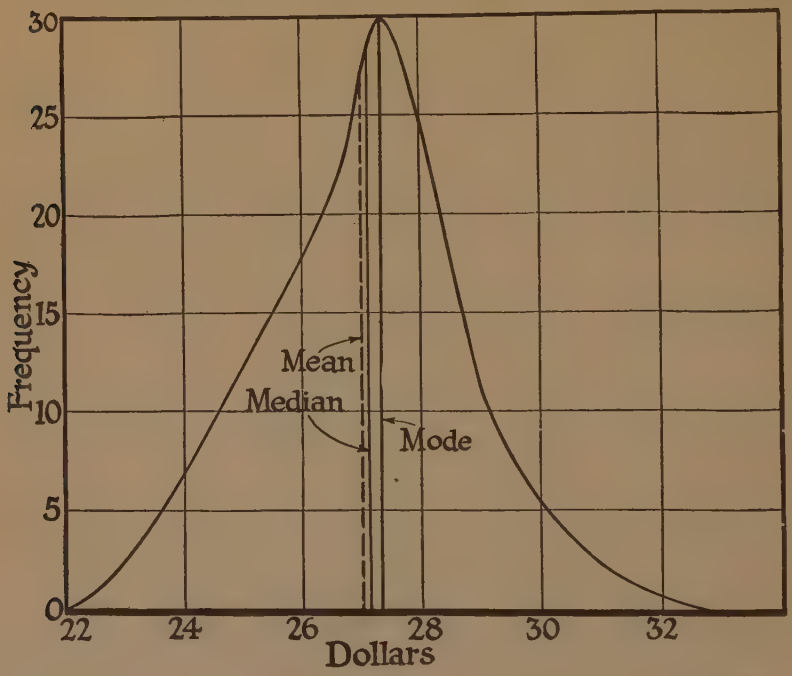


Fig. 45. — Distribution of Weekly Earnings of Employees. A Smoothed Frequency Curve, Showing the Relation between Mean, Median and Mode

In Fig. 46 the same data are represented by a cumulative frequency curve, based upon Table 27. The steepness of a cumulative frequency curve within any given interval depends upon the number of cases added within the corresponding interval on the horizontal scale. Thus the curve rises gradually at first, then more steeply, and tails off gradually at the upper extremity. The value of the mode, obviously, is the reading on the horizontal scale corresponding to the point of greatest steepness. This

is the point at which the increase of frequencies is greatest, the point of greatest concentration in the frequency distribution. The value of the mode may be approximated from a smoothed frequency curve by locating the point at which the slope is greatest (which is a point of inflection) and taking the corresponding reading on the x-scale. In the present case a value of approximately $27.50 is secured for the mode by this method.

TABLE 27

Cumulative Distribution of Wage-earners in a Manufacturing Establishment

(Classified on the Basis of Weekly Earnings.)

Weekly earnings	*Number earning stated amount (frequency)*
Less than $22.50	0
" " 23.00	1
" " 23.50	5
" " 24.00	8
" " 24.50	19
" " 25.00	29
" " 25.50	41
" " 26.00	56
" " 26.50	78
" " 27.00	98
" " 27.50	122
" " 28.00	152
" " 28.50	169
" " 29.00	186
" " 29.50	193
" " 30.00	199
" " 30.50	204
" " 31.00	208
" " 31.50	209
" " 32.00	209
" " 32.50	210

Values for the median, quartiles and deciles may also be secured graphically from the cumulative frequency curve. The smoothing of such a curve provides a quite satisfactory method of interpolation and, if the scale of the

diagram is sufficiently large, accurate values may be obtained by this method. Locate on the vertical scale (the scale of cumulative frequencies) a point distant from the base by $\frac{N}{2}$. If from this point a horizontal line be extended to the cumulative curve, the abscissa of the point of intersection will be the value of the median. This value may be easily determined by dropping a vertical line from the point of intersection to the x-axis. Figure 46 illustrates

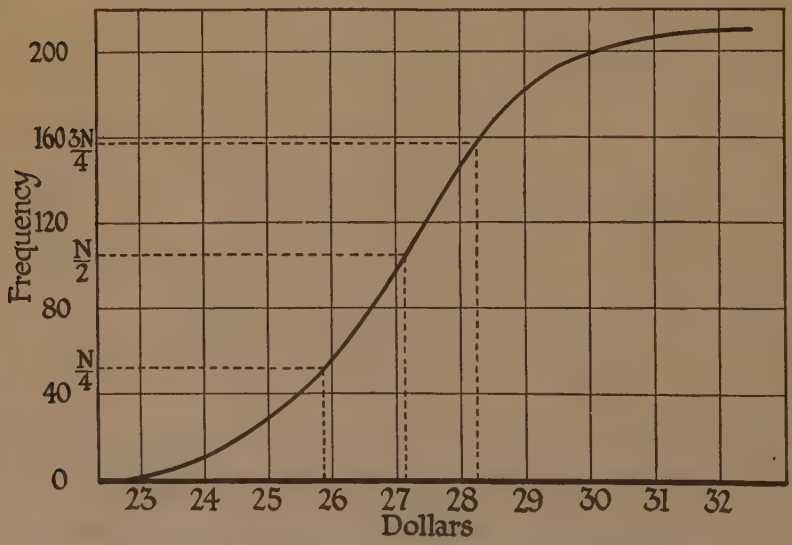


Fig. 46. — Cumulative Distribution of Weekly Earnings of Employees, Illustrating the Graphic Location of Median and Quartiles

the application of this method. A value of \$27.125 is secured for the median by this method. By direct interpolation a value of \$27.1458 is obtained. The quartiles may be located in precisely the same way, the vertical scale being divided into quarters and horizontal lines extended to the cumulative curve from the points thus located on the vertical scale.

For some purposes, particularly those which involve the

averaging of *rates* or *ratios* rather than quantities, none of the averages which have been described is suitable. The geometric and the harmonic means are types of averages which should be familiar because they are particularly appropriate for such purposes.

The Geometric Mean

The geometric mean is the nth root of the product of n measures; its value thus is represented by:

$$M_g = \sqrt[N]{a_1 \cdot a_2 \cdot a_3 \ . \ . \ . \ a_n}$$

The geometric mean of the numbers 2, 4, 8, is

$$\begin{aligned} M_g &= \sqrt[3]{2 \times 4 \times 8} \\ &= \sqrt[3]{64} \\ &= 4 \end{aligned}$$

It is obvious from the method of computation that if any of the measures in the series has a value of zero the geometric mean is zero.

The actual computation of the geometric mean is greatly facilitated by the use of logarithms. In this form

$$\text{Log } M_g = \frac{\log a_1 + \log a_2 + \log a_3 + \ . \ . \ . + \log a_n}{N}$$

The logarithm of the geometric mean is equal to the arithmetic mean of the logarithms of the individual measures.

When the measures, of which the geometric mean is desired, are to be weighted, the separate weights are introduced as exponents of the terms to which they apply. Thus if we represent the sum of the weights by N and the weights corresponding to the terms a_1, a_2, a_3 . . . a_n, respectively, by w_1, w_2, w_3 . . . w_n, the formula for the geometric mean is

$$M_g = \sqrt[N]{a_1^{w_1} \cdot a_2^{w_2} \cdot a_3^{w_3} \ . \ . \ . \ a_n^{w_n}}$$

This is equivalent to repeating each term a number of times, the number corresponding to the amount by which

it is weighted, which, of course, is precisely what is done in securing a weighted arithmetic mean. When logarithms are employed the formula for the weighted mean becomes

$$\text{Log } M_g = \frac{w_1 \log a_1 + w_2 \log a_2 + w_3 \log a_3 + \ldots + w_n \log a_n}{N}$$

The method of computing the geometric mean may be illustrated with reference to the following table, which shows the distribution of the prices of 115 dividend-paying preferred stocks listed on the New York Stock Exchange. The table is based upon closing prices on August 26, 1922.

TABLE 28

Computation of the Geometric Mean of Preferred Stock Prices

Class-interval	*m*	*f*	log *m*	*f* log *m*
$ 35–$ 44.9	$ 40	1	1.602060	1.602060
45– 54.9	50	6	1.698970	10.193820
55– 64.9	60	8	1.778151	14.225208
65– 74.9	70	5	1.845098	9.225490
75– 84.9	80	14	1.903090	26.643260
85– 94.9	90	22	1.954243	42.993346
95– 104.9	100	27	2.000000	54.000000
105– 114.9	110	18	2.041393	36.745074
115– 124.9	120	14	2.079181	29.108534
		115		224.736792

$$\text{Log } M_g = \frac{224.736792}{115}$$

$$\text{Log } M_g = 1.954233.$$

$$M_g = \$90.00$$

CHARACTERISTICS OF THE GEOMETRIC MEAN

The nature of the geometric mean may be understood by considering its relation to the terms it represents, as an average.

If the arithmetic mean of a series of measures replace each item in the series, the *sum* of the measures will remain unchanged. Thus, the sum of the numbers 2, 4, 8, is 14. The arithmetic mean of these three numbers is $4\frac{2}{3}$; if this value be inserted in the place of each of the three measures

the sum remains 14. It is characteristic of the geometric mean that the *product* of a series of measures will remain unchanged if the geometric mean of those measures replace each item in the series. Thus the product of 2, 4, 8, is 64. The geometric mean of the three numbers is 4; if this value replace each of the three measures the product remains 64.

Again, it is true of the arithmetic mean that the sum of the deviations of the items above the mean equals the sum of the deviations of the items below the mean (disregarding signs). The sums of the differences between the individual items and the mean are equal. In the case of the geometric mean the products of the corresponding ratios are equal. If the ratios of the geometric mean to the measures which it exceeds be multiplied together, the product will equal that secured by multiplying together the ratios to the geometric mean of the measures exceeding it in value. For example, the geometric mean of the numbers 3, 6, 8, 9, is 6. The following equation may be set up:

$$\frac{6}{3} \times \frac{6}{6} = \frac{8}{6} \times \frac{9}{6}$$

The last example brings out the most important characteristic of the geometric mean. It is a means of averaging ratios. Its chief use in the field of economic statistics has been in connection with index numbers of prices, where rates of change are of major importance. A rise in prices represented by the change from 50 to 100 is as important as a rise from 100 to 200. Yet this equivalence is not brought out by the arithmetic mean, which gives double weight to the change which involves an absolute difference of 100. An example frequently cited is that of two cases of price change, one a ten-fold increase, from 100 to 1000, the other a fall to one-tenth of the old price, from 100 to 10. The arithmetic mean of 1000 and 10 is 505, the geometric mean is $\sqrt{1000 \times 10}$, or 100. When the average is

of the latter type it is seen that the two equal ratios of change have balanced each other. The arithmetic mean, 505, is quite incorrect as a measure of average ratio of price change. This subject is discussed at greater length in the chapter on index numbers.

What has been said in an earlier section in regard to the advantages of logarithmic charting for certain purposes bears upon the use of the geometric mean. This average is sometimes called the logarithmic mean, as its logarithm is simply the arithmetic mean of the logarithms of the constituent measures. Wherever percentages of change are being averaged, where ratios rather than absolute differences are significant, the use of the geometric mean is advisable.

A problem involving the use of the geometric mean arises in computing the average rate of increase of any sum at compound interest. If p_o represent the principal at the beginning of the period, p_n the principal at the end of the period, r the rate of interest and n the number of years in the period, the sum to which p_o will amount at the end of the n years, if interest is compounded annually, is represented by the equation:

$$p_n = p_o(1 + r)^n$$

It follows from this that:

$$r = \sqrt[n]{\frac{p_n}{p_o}} - 1$$

Thus, if \$1000 at compound interest amounts to \$1600 at the end of 12 years, there has been an increase of 60%. The arithmetic mean is 5%, but this is not the rate at which the money increased. The true rate is:

$$r = \sqrt[12]{\frac{1600}{1000}} - 1$$

$$= \sqrt[12]{1.60} - 1$$

$$= 1.04 - 1$$

$$= .04, \text{ or } 4\%$$

Precisely the same problem arises whenever rates of increase or decrease are to be averaged. The use of the arithmetic mean gives an incorrect result.

The Geometric Mean as a Measure of Central Tendency

A question arises as to the type of frequency distribution the central tendency of which would be best represented by the geometric mean. This question has given rise to some interesting discussion. When the absolute measures, plotted on the arithmetic scale, give a fairly symmetrical distribution, the arithmetic mean is clearly preferable to the geometric mean. But when the absolute figures thus plotted give an asymmetrical frequency curve of such a type that the asymmetry would be removed and a symmetrical curve secured by plotting the logarithms of the measures, the geometric mean would appear to be preferable. Such a distribution would be one in which not the absolute deviations about the central tendency but the *relative* deviations, the deviations as ratios, were symmetrical. The arithmetic mean of the logarithms of the various measures (which value is, as has been shown, the logarithm of the geometric mean of the original measures) would be the best representative of the central tendency in such a distribution. The curve thus plotted would be symmetrical about the logarithm of the geometric mean. A frequency curve representing the logarithms of percentage changes in prices would tend to show this symmetry about the logarithm of the geometric mean of these changes. These percentage changes, as natural numbers, group themselves in an asymmetrical form, with the range of deviations above the arithmetic mean greatly exceeding the range below.[1] This arises, of course, from the fact that prices of given commodities may increase 1,000 per cent or more

[1] Cf. Figures 49 and 50.

from a given base, but cannot fall more than 100% from any given base. The section on index numbers contains a fuller discussion of this particular phase of the subject.[1]

A frequency curve of this type, based upon the logarithms of the measures included rather than upon the natural numbers, has been employed to advantage in plotting data relating to income distribution. When natural numbers are plotted, the range of income distribution is so large that it is physically impossible to prepare a chart which will reveal the characteristic features of all sections of the curve. The process of plotting on double logarithmic paper (which is, of course, equivalent to plotting the logarithms of both x's and y's) meets this difficulty, giving a true impression of the whole distribution and the relations between its parts, and, at the same time, brings out certain important features which are obscured in the natural scale chart. In particular, this device appears to smooth that part of the curve above the mode into a straight line, a fact which led Vilfredo Pareto, the first person to employ this method of representing income data, to enunciate what has been known as Pareto's Law concerning income distribution. An intensive study of the distribution of income in the United States has lead the staff of the National Bureau of Economic Research to call into question certain

[1] C. M. Walsh, in *The Problem of Estimation*, 35, lays down the following criteria for the use of averages:

(*a*) When there are no conceivable or assignable upper or lower limits to the values of the terms in a series, the arithmetic average should be employed.

(*b*) When there is a definite lower limit at or above zero and no upper conceivable or assignable limit, the geometric average should be employed. Because this is true of price changes Walsh believes the geometric average to be the correct one to use in making index numbers of prices.

(*c*) When in practice, or in the nature of things, certain upper and lower limits are found to exist and the above criteria cannot be employed, a study of the actual dispersion of the data is necessary. In this case, if the mode is found nearer to the arithmetic average, that average should be employed; if the mode is found nearer to the geometric average, that average should be used.

conclusions drawn from Pareto's generalizations, though the value of the double logarithmic scale for the presentation of income data has been recognized. An interesting discussion of this type of frequency curve is found in the reports on *Income in the United States*, prepared by the National Bureau of Economic Research.

The Harmonic Mean

The harmonic mean is a type of average capable of application only within a restricted field, but which should be employed to avoid error in handling certain types of data. It must be used in the averaging of time rates and it has certain distinctive advantages in the manipulation of certain other materials. The following example will illustrate the method of employing this average:

An automobile is driven four miles at the rate of twenty miles per hour and four miles at the rate of thirty miles per hour; the average rate of speed is required. The arithmetic mean of the two rates is 25 miles per hour, but this is an incorrect solution. The machine is driven 12 minutes at the rate of 20 miles per hour and 8 minutes at the rate of 30 miles. Thus in 20 minutes 8 miles are covered, the average rate being 24 miles per hour. This is equivalent to taking a weighted average of 20 and 30, the former weighted by 12 and the latter by 8. The same result may be secured by taking the harmonic mean of the two rates directly. *The harmonic mean of a series of measures is the reciprocal of the arithmetic mean of the reciprocals of the individual measures.* Thus if we represent a series of rates to be averaged by r_1, r_2 . . . r_n, the formula for the harmonic mean, H, is

$$\frac{1}{H} = \frac{\frac{1}{r_1} + \frac{1}{r_2} + \frac{1}{r_3} + \ldots + \frac{1}{r_n}}{N}.$$

Using the figures just quoted:

$$\frac{1}{H} = \frac{\frac{1}{20} + \frac{1}{30}}{2}$$

$$= \frac{5}{120}$$

$$H = 24.$$

When the harmonic mean of two numbers, a and b, is required, it is simpler to work from the form $H = \frac{2ab}{a+b}$; for three numbers, a, b and c, the form $H = \frac{3abc}{ab + ac + bc}$ may be used. These are equivalent to the first formula given. The computation of the harmonic mean of a series of magnitudes is greatly facilitated by the use of prepared tables of reciprocals.[1]

The use of the harmonic mean in handling economic data may be illustrated by an example from the price field. If a given commodity is quoted at "so many per dollar," and it is desired to average several such quotations, the use of the simple arithmetic mean may give erroneous results. Given the quotations "four for a dollar," "five for a dollar" and "twenty for a dollar," the average number per dollar (i.e. the average price) is required. The arithmetic average of the figures as given (4, 5 and 20) is $9\frac{2}{3}$ which would appear to be the average number sold per dollar, an average price of 10.34 cents apiece. But the original quotations are equivalent to prices of 25¢ apiece, 20¢ apiece and 5¢ apiece; the arithmetic average of these prices is $16\frac{2}{3}$¢, which, expressed in the original form, gives an average of 6 per dollar. Assuming that it is desired to give equal weight to each quotation, the latter is the correct figure.

[1] *Barlow's Tables of Squares, Cubes, Square Roots, Cube Roots, Reciprocals* (N. Y., Spar and Chamberlain, 1919) constitute a very useful compilation.

The arithmetic mean of the quotations in the "so many per dollar" form is a weighted average, greater weight being given to quotations involving a large number of commodity units. The correct result may be secured from the original quotations, however, by the use of the harmonic mean. The harmonic mean of 4, 5, and 20, is 6, the average number per dollar.

Relations Between Different Averages

When different averages are located or computed for a given series of magnitudes, certain relationships between them are found to prevail.

1. The arithmetic mean, the median and the mode coincide in a symmetrical distribution.
2. In a moderately asymmetrical distribution the median lies between the mean and the mode, approximately one third of the distance along the scale from the former towards the latter. Hence, for this type of distribution there is an approximation to the following relationship:

$$Mo = M - 3(M - Md).$$

3. The arithmetic mean of any series of magnitudes is greater than their geometric mean.
4. The geometric mean of any series of magnitudes is greater than their harmonic mean. The only exception to the last two rules is found when all the measures in the series are equal, in which case arithmetic mean, geometric mean and harmonic mean are equal.
5. The geometric mean of any two terms is equal to the geometric mean of the harmonic and arithmetic means of those terms. Thus if the terms be 2 and 8, the harmonic mean is $3\frac{1}{5}$, the geometric mean 4, and the arithmetic mean 5. But 4 is also the geometric mean of $3\frac{1}{5}$ and 5. This relationship does not hold when the series includes more than two terms.
6. When the dispersion of data follows the arithmetic law, the mode and median will generally be found closer to the arithmetic than to the geometric average. When the

dispersion follows the geometric law the mode and median will generally be found closer to the geometric than to the arithmetic average.

Characteristic Features of the Chief Averages

The arithmetic mean

1. The value of the arithmetic mean is affected by every measure in the series. For certain purposes it is too much affected by extreme deviations from the average.
2. The arithmetic mean is easily calculated, and is determinate in every case.
3. The arithmetic mean is a computed average, and hence is capable of algebraic manipulation.

The median

1. The value of the median is not affected by the magnitude of extreme deviations from the average.
2. The median may be located when the items in a series are not capable of quantitative measurement.
3. The median may be located when the data are incomplete, provided that the number and general location of all the cases be known, and that accurate information be available concerning the measures near the center of the distribution.
4. The median is not so well adapted to algebraic manipulation as are the arithmetic, geometric and harmonic means.

The mode

1. The value of the mode is not affected by the magnitude of extreme deviations from the average.
2. The approximate mode is easy to locate but the determination of the true mode requires extended calculation.
3. The mode has no significance unless the distribution include a large number of measures and possess a distinct central tendency.
4. The mode is the average most typical of the distribution, being located at the point of greatest concentration.
5. The mode is not capable of algebraic manipulation.

The geometric mean

1. The geometric mean gives less weight to extreme deviations than does the arithmetic mean.
2. The geometric mean is strictly determinate in every case.
3. The geometric mean is the form of average to be used when rates of change or ratios between measures are to be averaged, as equal weight is given to equal ratios of change. It is particularly well adapted to the averaging of ratios of price change.
4. The geometric mean is capable of algebraic manipulation.

The harmonic mean

1. The harmonic mean is adapted to the averaging of time rates and certain similar terms. It has been employed in the field of economic statistics in the manipulation of price data.
2. The labor of computing the harmonic mean and its unfamiliarity detract from its usefulness in ordinary statistical analysis.
3. The harmonic mean is capable of algebraic manipulation.

This summary has been designed to show that each type of average has its own particular field of usefulness. Each one is best for certain purposes and under certain conditions. The characteristics and limitations of each one should be understood in order that it may be appropriately employed. A complete description of a frequency distribution frequently calls for the determination of two or three of the chief averages, as well as other statistical measurements. The arithmetic mean is perhaps the most useful single average. The simplicity of its computation, the possibility of employing it in algebraic calculations and the fact that its meaning is perfectly definite and familiar make it highly serviceable in statistical work. Its sphere of usefulness is not universal, however, and it should only be employed when the given conditions render it suitable. A fuller appreciation of the distinctive virtues of the geometric mean is leading to a wider employment of that

measure in many types of statistical work. A discriminating use of averages is essential to sound statistical analysis.

REFERENCES

Bowley, A. L. *Elements of Statistics* (82–109, 138–139).
Chaddock, R. E. *Principles and Methods of Statistics* (Chaps. VI, VII, VIII).
Jones, D. C. *A First Course in Statistics* (22–41).
Kelley, Truman L. *Statistical Method* (44–68).
King, W. I. *Elements of Statistical Method* (121–140).
Pearl, Raymond. *Medical Biometry and Statistics* (264–272).
Rugg, H. O. *Statistical Methods Applied to Education* (97–148).
Secrist, Horace. *Introduction to Statistical Methods* (234–292).
Walsh, C. M. *The Problem of Estimation* (1–67).
Yule, G. U. *An Introduction to the Theory of Statistics* (106–132).
Zizek, Franz. *Statistical Averages* (82–109).

CHAPTER V

DESCRIPTION OF THE FREQUENCY DISTRIBUTION: MEASURES OF VARIATION AND SKEWNESS

In the preceding chapters we have been concerned, first, with methods of reducing a mass of quantitative data to a form in which the characteristics of the mass as a whole may be readily determined and, in the second place, with methods of describing the assembled data. The first object is accomplished with the formation of a frequency distribution. The second is partially accomplished when there has been obtained a single significant value in the form of an average which represents the central tendency of the distribution. But any average, by itself, fails to give a complete description of a frequency distribution. Three other values are needed before the chief characteristics of a given distribution have been measured, and comparison with other distributions is possible. The first of these is a measure of the degree to which the items included in the original distribution depart or *vary* from the central value, the degree of "*scatter*," *variation* or *dispersion*. The second is a measure of the degree of symmetry of the distribution, of the balance or lack of balance on the two sides of the central value. The third is a measure of *kurtosis*, of the degree to which there is a bunching of cases at the modal value. The present chapter deals with various measures of variation and skewness. The method of measuring kurtosis is referred to at a later point.

Nature and Significance of Variation

The fact of variation in collections of quantitative data has been pointed out in earlier sections and the bearing of

this fact upon the work of the statistician indicated. Practically every collection of quantitative data, consisting of measurements from the social, biological or economic field, is characterized by variation, by quantitative differences between the individual units. And this fact of variation is as important as the fact of family resemblance. Biological variation has been a fundamental factor in the evolutionary process. No measurement of a physical characteristic of a racial group, such as height, is complete without an accompanying measure of the average variation in the group in this respect. The average income in a country is perhaps of less significance than the variation in income, the differences between the incomes received by different social classes. Price variations interrupt the normal functioning of the economic system, causing hardship to some and giving unearned profits to others, because the various elements in the price system are unequally affected. It is not the general change in the price level but the differences between price changes which cause trouble.

An average, by itself, has little significance unless the degree of variation in the given frequency distribution is known. If the variation is so great that there is no pronounced central tendency an average has no significance. With a decrease in the degree of variation an average becomes increasingly significant. Whether a single frequency distribution is being described, therefore, or comparison is being made with other distributions, a measure of central tendency must be supplemented by a measure of variation.

Measures of Absolute Variation

Variation may be expressed in terms of the units of measurement employed for the original data, or may be expressed as an abstract figure, such as a percentage, which is independent of the original units. When the original units are employed *absolute variability* is measured;

when an abstract figure is secured we have a measure of *relative variability*, more suitable for comparison than the former type. Measures of absolute variability are first considered.

The Range

A rough measure of variation is afforded by the *range*, which is the absolute difference between the value of the smallest item and the value of the greatest item included in the distribution. Table 30 shows the distribution of London-New York monthly exchange rates during the period 1882–1913. The smallest item among the original figures included in the table is \$4.83; the greatest is \$4.908. The range, therefore, is \$4.908–\$4.83, or \$.078. A distance on the scale equal to \$.078 will include every item. If the original data were not to be had the range could be approximated from the frequency table. It would be the difference between the lower limit of the class at the lower extreme of the distribution, and the upper limit of the class at the upper extreme, or \$.085 in the present case.

The value of the range, it is obvious, depends upon the values of the two extreme cases only. A single abnormal item would change its value materially. Because it is erratic and is likely to be unrepresentative of the true distribution of items, it is seldom used in statistical work. The range is frequently employed as a measure of stock market fluctuations, though its adequacy for this purpose may be questioned.

The Mean Deviation

A more accurate measure of the dispersion of items about a central value is afforded by the simple device of measuring the deviation of each item from this central value and averaging these deviations. The following simple example illustrates the method of computation:

TABLE 29

Computation of Mean Deviation

m	f	d
3	1	6
6	1	3
9	1	0
12	1	3
15	1	6
		18

$M = 9$

$M.D. = \frac{18}{5} = 3.6$

The average (the mean and median coincide in this case) is 9. The deviations are totaled, taking no account of algebraic signs, and the total divided by the number of items. This procedure is described by the expression

$$M.D. = \frac{\Sigma d}{N}.$$

In general terms, the *mean deviation* of a series of magnitudes is the arithmetic mean of their deviations from an average value (either mean or median). In the process of summation and averaging the algebraic signs of the deviations are disregarded. In practice it makes little difference whether deviations be measured from the mean or the median. Theoretically the latter should be chosen, for the value of the mean deviation is least when the median is the base.

Table 30 illustrates the computation of the mean deviation when the data are grouped in a frequency distribution. In the computation of the arithmetic mean of the items of a frequency distribution the assumption is made, it will be remembered, that all the cases in a given class are concentrated at the mid-point of that class; in other words, that the mid-value of the class may be taken as the value of each of the measures included. The same assumption is made in the computation of the mean deviation.

TABLE 30

Computation of Mean Deviation

London-New York Exchange Rates, 1882–1913

(The table is based upon rates prevailing at the beginning of each month)

Class-interval	*Mid-point*	*Frequency*	*Deviation from arbitrary origin*		
	m	*f*	*d'*	*fd'*	
$4.8275–$4.8324	$4.830	1	$.040	$ .040	Arbitrary origin = $4.8700
4.8325– 4.8374	4.835	6	.035	.210	
4.8375– 4.8424	4.840	11	.030	.330	Median = 4.8721
4.8425– 4.8474	4.845	21	.025	.525	
4.8475– 4.8524	4.850	23	.020	.460	Difference = $ .0021
4.8525– 4.8574	4.855	24	.015	.360	
4.8575– 4.8624	4.860	25	.010	.250	196 cases too small by
4.8625– 4.8674	4.865	40	.005	.200	$.0021
4.8675– 4.8724	4.870	45	....	.....	188 cases too large by
4.8725– 4.8774	4.875	49	.005	.245	$.0021
4.8775– 4.8824	4.880	35	.010	.350	Net result: 8 cases too
4.8825– 4.8874	4.885	45	.015	.675	small by $.0021
4.8875– 4.8924	4.890	33	.020	.660	Sum of deviations from
4.8925– 4.8974	4.895	16	.025	.400	arbitrary origin = $5.020
4.8975– 4.9024	4.900	8	.030	.240	Correction (8 times
4.9025– 4.9074	4.905	1	.035	.035	$.0021) = $.0168
4.9075– 4.9124	4.910	1	.040	.040	Sum of deviations from
					median = $5.020 + .0168
		384		$5.020	= $5.0368
					$M.D. = \frac{\$5.0368}{384}$
					= $.01312

The mean deviation might be computed from this table by the method followed in Table 29. The deviation of the mid-point of each class from the median would be computed, multiplied by the number of cases in that class, and the total divided by the number of items. Thus the median of the exchange rates is $4.8721; the deviation of the item in the first class is $.0421, of each of the six items in the second class $.0371, etc. But the handling of these

true deviations is laborious, since fractional quantities are involved. It is simpler to measure the deviations from some arbitrary value, sum these deviations from the assumed median (or mean), and correct this total by an amount equal to the error caused by the measurement of deviations from a value other than the true average. This is the method followed in Table 30.

The same method would be employed in computing the mean deviation from the arithmetic mean. In either case the computations might be simplified still further by measuring deviations in class-interval units, as in some of the calculations in the preceding chapter.

In Table 30 deviations have been measured not from \$4.8721, the value of the median, but from \$4.870, which is selected arbitrarily as the origin. The sum of the deviations from this arbitrary origin, regardless of signs, is \$5.020. By how much does this differ from the sum of the deviations from the median? For each of the 45 cases in the class having as mid-point the arbitrary origin, \$4.870, the value of d' (the deviation from the arbitrary origin) is zero. They deviate from the median, however, by the difference between \$4.870 and \$4.8721, or \$.0021. In this case the measured deviations are too small by \$.0021. The same error is made in measuring the deviation of each item in all the lower classes; in all there are 196 cases in which the deviation from the arbitrary origin is less by \$.0021 than the deviation from the median. But for all the observations falling in classes above that in which the arbitrary origin lies, the error is in the opposite direction. Thus, for the 49 items in the class next above that in which the origin falls, the deviation from the arbitrary origin is \$.005, while the deviation from the median is \$.0029; the deviations as measured are too large by \$.0021. In all, there are 188 cases in which the error is in this direction. The net result, therefore, is that 8 are too small by \$.0021. The sum of the deviations, \$5.020 is, therefore, too small

by $8 \times \$.0021$, or $\$.0168$. The sum of the deviations from the median is thus $\$5.0368$, and the mean deviation is $\$.01312$.

This method may be summarized in the following formula:

$$M.D. = \frac{\Sigma(fd') + (N_s - N_l)c}{N}$$

in which

N_s = Number of cases for which deviations measured from arbitrary origin are smaller than deviations measured from median (or mean).

N_l = Number of cases for which deviations measured from arbitrary origin are larger than deviations measured from median (or mean).

c = Difference between arbitrary origin and the median (or mean).

For the application of this formula it is necessary that the arbitrary origin and the median (or mean) be located within the same class-interval.[1]

The following outline indicates the steps to be taken in the computation of the mean deviation:

1. Determine the value of the median.
2. Taking the mid-point of the class containing the median as arbitrary origin, measure the deviations from this origin of the items in each class. Multiply these deviations by the class-frequencies and secure the total of the products without regard to sign.
3. Count the number of cases in which the deviations from the arbitrary origin are greater than the deviations from the median, and the number of cases in which they are less than the deviations from the median. Get the difference between these two figures; multiply by this quantity the

[1] The above calculations proceed on the assumption that all the cases in the class containing the median may be treated as though they were concentrated at the mid-point of that class. A slight gain in accuracy might be made by assuming a uniform distribution throughout this class-interval and modifying the method accordingly. (Cf. *Handbook of Mathematical Statistics*, edited by H. L. Rietz, 29–31, for an exposition of this method.)

difference between the median and the arbitrary origin. This is the amount by which the sum of the deviations from the arbitrary origin differs from the sum of the deviations from the median. Apply the above correction to the sum of the deviations from the arbitrary origin.

4. Divide the corrected sum of the deviations by the number of items to get the mean deviation from the median.

(The mean deviation from the mean may be computed by an identical process.)

The Standard Deviation

The process of calculating the mean deviation is algebraically illogical because algebraic signs are disregarded. In the computation of the standard deviation this error is avoided and a measure of more precise mathematical significance is secured. The conventional symbol for the standard deviation is the Greek letter *sigma*, σ.

In computing this measure the deviations of the individual items from the arithmetic mean are squared, totaled, the mean of the squared deviations obtained, and the square root of this mean extracted. The *standard deviation* is, thus, the square root of the mean of the squared deviations. This measure is also termed the *root-mean-square deviation*, a useful name because it describes in full the method of calculation. The deviations are always measured from the arithmetic mean, as the value of the measure is a minimum under these conditions. A simple example will illustrate the process.

Table 31

Computation of Standard Deviation

m	f	d	d^2	
3	1	− 6	36	$M = 9$
6	1	− 3	9	
9	1	0	0	$\sigma = \sqrt{\frac{90}{5}}$
12	1	+ 3	9	
15	1	+ 6	36	$= \sqrt{18}$
	5		90	$\sigma = 4.24$

When the standard deviation is computed from ungrouped data, as in this case, the formula is

$$\sigma = \sqrt{\frac{\Sigma d^2}{N}}.$$

When the items are grouped in a frequency distribution the task of computation is a little more complicated. The measurement of deviations from an arbitrary origin is essential in this case, as it greatly simplifies the calculations. The general formula for the standard deviation is

$$\sigma = \sqrt{\frac{\Sigma fd^2}{N}}$$

where f represents the class-frequencies, d the deviations from the arithmetic mean and N the number of cases included. It follows that

$$\sigma^2 = \frac{\Sigma fd^2}{N}.$$

If a deviation from an arbitrary origin be represented by d' and the root-mean-square deviation from this origin be represented by s, we have

$$s^2 = \frac{\Sigma f(d')^2}{N}.$$

The root-mean-square deviation from the mean (σ) is less than the root-mean-square deviation from any other point on the scale. Hence s^2 is greater than σ^2. We may represent by c the difference between the true mean and the arbitrary origin. It may be readily established [1] that

$$\sigma^2 = s^2 - c^2.$$

[1] For $\sigma^2 = \frac{\Sigma d^2}{N}$ but $\Sigma d = 0$

$s^2 = \frac{\Sigma (d')^2}{N}$ $\quad \therefore \Sigma (d')^2 = \Sigma d^2 + Nc^2$

$d' = d + c$ $\quad \frac{\Sigma (d')^2}{N} = \frac{\Sigma d^2}{N} + c^2$

$(d')^2 = d^2 + 2cd + c^2$ $\quad s^2 = \sigma^2 + c^2$

$\Sigma (d')^2 = \Sigma d^2 + 2c\Sigma d + Nc^2$ $\quad \sigma^2 = s^2 - c^2$

Cf. Yule, *Introduction to the Theory of Statistics*, 134.

The value of the standard deviation may be most easily determined, therefore, by computing s^2 and c^2. The operations involved are illustrated in detail in Table 32, showing the distribution of London-Paris exchange rates over a period of 384 months.

TABLE 32

Computation of Standard Deviation

London-Paris Exchange Rates, 1882–1913

(The table is based upon rates prevailing at the beginning of each month, as quoted in the *Economist*)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Class-interval (francs)	*Mid-point (francs)*	*Frequency*	*Deviation from arbitrary origin*				
	m	f	d'	fd'	$f(d')^2$	$(d'+1)^2$	$f(d'+1)^2$
25.07–25.089	25.08	1	− 8	− 8	64	49	49
25.09–25.109	25.10	4	− 7	− 28	196	36	144
25.11–25.129	25.12	14	− 6	− 84	504	25	350
25.13–25.149	25.14	20	− 5	− 100	500	16	320
25.15–25.169	25.16	45	− 4	− 180	720	9	405
25.17–25.189	25.18	60	− 3	− 180	540	4	240
25.19–25.209	25.20	40	− 2	− 80	160	1	40
25.21–25.229	25.22	43	− 1	− 43	43	0	
25.23–25.249	25.24	42	0			1	42
25.25–25.269	25.26	32	1	32	32	4	128
25.27–25.289	25.28	26	2	52	104	9	234
25.29–25.309	25.30	21	3	63	189	16	336
25.31–25.329	25.32	20	4	80	320	25	500
25.33–25.349	25.34	4	5	20	100	36	144
25.35–25.369	25.36	6	6	36	216	49	294
25.37–25.389	25.38	2	7	14	98	64	128
25.39–25.409	25.40	2	8	16	128	81	162
25.41–25.429	25.42	2	9	18	162	100	200
		384		− 372	4,076		3,716

$N = 384$

class-interval = .02 francs

c (in class-interval units) $= \frac{-372}{384} = -.969$

c^2 (in class-interval units) = .9390

s^2 (in class-interval units) $= \dfrac{\Sigma f(d')^2}{N} = \dfrac{4{,}076}{384} = 10.6146$

σ^2 (in class-interval units) $= s^2 - c^2 = 10.6146 - .9390 = 9.6756$

σ (in class-interval units) $= 3.11$

σ (in original units) $= 3.11 \times .02 = .0622$

The entire calculation, it will be noted, is carried through in terms of class-interval units, the result being reduced to the original units in the final operation. In computing c, the difference between the true mean and the arbitrary origin, the algebraic sum of the deviations is divided by the number of cases. The arithmetic mean could be determined by reducing c to original units and adding this value (algebraically) to the value of the arbitrary quantity selected as origin, but this is not an essential step. The actual value of the mean need not be known in the computation of the standard deviation.

A check upon the accuracy of the calculations (the *Charlier check*[1]) is afforded by the figures in cols. (7) and (8) of Table 32. If deviations be measured, not from the arbitrary origin employed in computing the standard deviation, but from an origin one class-interval below, we secure a set of values equal to $d' + 1$. The squares of these values are given in col. (7). Multiplying by the corresponding frequencies we have the quantities recorded in col. (8), the sum of which is 3,716. This total stands in a definite relationship to the values secured in computing the standard deviation. For

$$\Sigma f(\mathrm{d}' + 1)^2 = \Sigma f[(d')^2 + 2d' + 1]$$
$$= \Sigma f(d')^2 + 2\Sigma f d' + \Sigma f$$

or
$$\Sigma f(d' + 1)^2 = \Sigma f(d')^2 + 2\Sigma f d' + N$$

Inserting in this last equation the values secured from the calculations shown in Table 32, we obtain this check:

$$3{,}716 = 4{,}076 + 2(-372) + 384$$
$$= 3{,}716$$

[1] Cf. *Vorlesungen Über Die Grundzüge Der Mathematischen Statistik*, C. V. L. Charlier, 19.

The following is a summary of the steps in the process of computing the standard deviation of items grouped in a frequency distribution:

1. Select as arbitrary origin the mid-point of a class near the center of the distribution.
2. Measure the deviations from this point of the items in each class, in class-interval units. Multiply the deviations by the corresponding class-frequencies.
3. Divide the algebraic sum of the deviations by N. This gives c, in class-interval units. Compute c^2.
4. Square the deviations and multiply by the corresponding class frequencies.
5. Divide the sum of the squared deviations by N. This gives s^2, in class-interval units.
6. From the formula, $\sigma^2 = s^2 - c^2$, compute σ^2. Extract the square root of this value, securing σ in class-interval units.
7. Multiply σ, as thus computed, by the class-interval. The result is σ in the original units of measurement.

Certain of the characteristics of the standard deviation and its relation to other measures of dispersion are described in a later section.[1]

The Quartile Deviation

In the chapter on averages methods of locating the quartiles and deciles were described. The former are those points on the scale of values, along which the items of a given distribution lie, which divide the total number of items into four equal groups. The deciles are those points dividing the total number of items into ten equal groups. The degree and character of the variation in a frequency distribution may be accurately described if the location of the quartiles and deciles is shown. Such knowledge, however, while helpful in giving a picture of the distribution, is not as useful for purposes of concise description and

[1] A correction to be applied to the standard deviation in certain cases (Sheppard's correction) is described in Chapter XV.

comparison as knowledge of the values of the mean deviation or the standard deviation. The significance of a single measure is more readily grasped than is the meaning of a number of inter-related values. Such a measure of variation may be computed from the quartiles, however. With regard to ease of calculation and immediate significance this *quartile deviation* has distinct merits.

Within the range between the two quartiles, of course, one half of all the measures are included. The greater the concentration the smaller this interval, hence a fairly accurate measure of dispersion may be obtained from the relationship between these two quartiles. The quartile deviation is the *semi-interquartile range*, half the distance along the scale between the first and third quartiles. Thus if $Q.D.$ represent the quartile deviation, Q_1 the first quartile and Q_3 the third quartile,

$$Q.D. = \frac{Q_3 - Q_1}{2}.$$

If the value of a point on the scale half-way between the first and third quartiles is represented by K, one half of all the measures in a frequency distribution will fall within the range $K \pm Q.D.$ For the preceding data, London-Paris exchange rates, we have

$$\begin{aligned} Q_3 &= 25.262 \\ Q_1 &= 25.174 \\ Q.D. &= \frac{25.262 - 25.174}{2} \\ &= .044 \\ K &= 25.174 + .044 \\ &= 25.218. \end{aligned}$$

Thus one half of all the measures lie within the range $25.218 \pm .044$. This statement, together with a statement of the average value of exchange rates (mean, median or mode), constitutes a useful description of the distribution. In a perfectly symmetrical distribution the value of K will

coincide with the value of the median (that is, the median will lie half-way along the scale from Q_1 to Q_3). The distribution of exchange rates is asymmetrical, the value of the median being 25.214, as compared with the value of 25.218 for K.

The Probable Error

In studying the results of astronomical and other physical measurements it has been found that the values secured by different observers for the same constant quantity vary. These varying results, however, are distributed in a certain definite way, and when plotted give a curve similar to the normal curve of error. In such cases there is an immediate and obvious need of some measure of variation which may be used as an index of the reliability of given results. If the results secured by different investigators, or by the same investigator at different times, vary widely they can not be accepted as reliable, while the reverse is true if the variation is slight. The measure of dispersion which has been generally employed in such cases is termed the *probable error*. The probable error is that amount which, in a given case, is exceeded by the errors of one half the observations. Since the most probable value of a given series of observations is their arithmetic mean, the probable error is always measured from the mean. The name of this measure derives from the fact that the probability that a given observation will vary from the mean of all the observations by an amount greater than the probable error is exactly $\frac{1}{2}$. It follows that, when the observations are arranged in the form of a frequency distribution, an amount equal to the probable error laid off on each side of the arithmetic mean will include one half of the total number of cases.

This measure of variation has been employed in fields other than that in which it was originally applied, fields in which the original name of probable error is somewhat

misleading. In such cases it is perhaps better to think of it as the *probable deviation*, that distance from the mean which will be exceeded by one half of the total deviations.

The probable error is a measure of dispersion which is fully significant only when it applies to a distribution following the normal law of error. In such cases it has a definite and precise meaning. This is not so when it is applied to skewed distributions, and its use in such cases is not advisable. The quartile deviation, the value of which is equal to that of the probable error in a normal distribution, has a more direct significance than the probable error in the description of abnormal distributions, and should be employed in such cases. In a later section the use of the probable error as a measure of the reliability of statistical results is more fully explained.

The value of the probable error in a given case, assuming a normal distribution to prevail, may be determined from the value of the standard deviation, for there is a constant relationship between these two. This is expressed by the formula: $P.E. = 0.6745\,\sigma$.

Relations Between Different Measures of Variation

An understanding of the significance of the various measures of dispersion described above may be facilitated by a general comparison and a summary statement of the relations holding between them.

1. The *range* is a distance along the scale within which all the observations lie.
2. The *quartile deviation* or *semi-interquartile range* is a distance along the scale which, when laid off on each side of the point midway between the two quartiles, includes one half of the total number of observations.
3. The *mean deviation* from the mean, in a symmetrical or slightly skewed distribution, is equal to about $\frac{4}{5}$ of the standard deviation. A range of $7\frac{1}{2}$ times the mean

deviation, centering at the mean, will include approximately 99% of all the cases.

4. When a distance equal to the *standard deviation* is laid off on each side of the mean, in a normal or only slightly skewed distribution, about two thirds of all the cases will be included. (In the normal distribution exactly 68.26% of the observations will be included.) When a distance equal to twice the standard deviation is laid off on each side of the mean approximately 95% of the cases will be included (exactly 95.46% in a normal distribution). When a distance equal to three times the standard deviation is laid off on each side of the mean about 99% of all the observations will be included (exactly 99.73% in a normal distribution). This general rule that a range of six times the standard deviation, centering at the mean, will include about 99% of all the measures furnishes a useful check upon calculations.

A study of Fig. 43 may help to make clear the significance of the standard deviation in a normal distribution.

5. The *probable error*, in a normal distribution, is equal to 0.6745σ. A range of twice the probable error, centering at the mean, will include 50% of all the observations. A range of eight times the probable error, centering at the mean, will include approximately 99% of all the observations.

Characteristic Features of the Chief Measures of Variation

The Range

1. The range is easily calculated and its significance is readily understood. As a rough measure of the degree of variation the range is useful.
2. The value of the range is determined by the values of the two extreme cases. It is thus a highly unstable measure, the value of which may be greatly changed by the addition or withdrawal of a single figure.
3. This measure gives no indication of the character of the distribution within the two extreme observations.

The quartile deviation

1. The quartile deviation is a measure of dispersion which is easily computed and readily understood. It is superior to the range as a rough measure of variation.
2. The quartile deviation is not a measure of the variation from any specific average.
3. This measure is not affected by the distribution of the items within the first or third quartiles, or by the distribution outside the quartiles. The values of the quartile deviation might be the same for two quite dissimilar distributions, provided the quartiles happened to coincide. Because it is not affected by the deviations of individual items it cannot be accepted as an accurate measure of variation.
4. The quartile deviation is not capable of algebraic treatment.

The mean deviation

1. The mean deviation is affected by the value of every observation. As the *average difference* between the individual items and the median (or mean) of the distribution it has a precise significance.
2. The mean deviation is less affected by extreme deviations than the standard deviation.
3. Mathematically, the mean deviation is not as logical or as convenient a measure of dispersion as the standard deviation.

The standard deviation

1. The standard deviation is affected by the value of every observation.
2. The process of squaring the deviations before adding avoids the algebraic fallacy of disregarding signs.
3. The standard deviation has a definite mathematical meaning and is perfectly adapted to algebraic treatment.
4. The standard deviation is, in general, less affected by fluctuations of sampling than the other measures of dispersion.
5. The normal curve of error has been analyzed in terms of the standard deviation. The information thus obtained has increased greatly the utility of the standard deviation.

The probable error

1. The probable error has a definite meaning in the case of a distribution following the normal law. It has not this precise meaning for other distributions, and should not be employed in describing them.
2. For distributions to which it is adapted, the probable error is an extremely useful measure. Its most important use is as an index of the reliability of quantitative measurements.
3. The definite relationship between the probable error and the standard deviation, for a normal distribution, permits the value of the probable error to be readily determined.

All the measures of variation described above may be utilized for particular purposes. The standard deviation, however, is the best general measure and should be employed in all cases where a high degree of accuracy is required. The probable error is, in effect, merely a fractional part of the standard deviation, with a definite but restricted field of usefulness.

The Measurement of Relative Variation

We have been dealing in the preceding section with absolute variability. The various measures of dispersion secured by the methods outlined describe the variability of the data in terms of absolute units of measurement. The standard deviation of London-Paris exchange rates is in francs, the standard deviation of pig iron production in tons, etc. If the object in a given case is the description of a single frequency distribution it is desirable that the original unit be employed throughout, but if measures of variation of two different distributions are to be compared, difficulties are encountered. This is clear if the units are unlike, but even if the units are identical the same difficulty arises. Thus measures of variation in the weights of dogs and in the weights of horses might

have been computed, both in pounds. Because the standard deviation of horse weights is greater than the standard deviation of dog weights, it does not follow that the degree of variability is greater in the former case. A measure of absolute variation is significant only in relation to the average from which the deviations are measured. Its use, apart from this average, is meaningless. For comparison, therefore, it must be reduced to a relative form, and the obvious procedure is to express a given measure of variation as a percentage of the average from which the deviations have been measured. The quantity thus becomes an abstract number, a measure of the relative variability of the given observations, and may be compared with similar terms computed from other distributions.

The Coefficient of Variation

The measure of relative variation most commonly employed is that developed by Pearson, termed the *coefficient of variation*, and represented by the letter V. It is simply the standard deviation as a percentage of the arithmetic mean. Thus

$$V = \frac{\sigma}{M} \times 100.$$

Applying this formula to the results secured from the analysis of London-Paris exchange rates, we have

$$V = \frac{.0622}{25.2206} \times 100.$$

$$= .25\%.$$

The variability of London-New York exchange rates during the same period may be computed from the data of Table 30. The coefficient of variation for this series has a value of .33%. Fluctuations in the London-New York rate were significantly more pronounced than in the London-Paris rate during this period.

An index of variability similar to this coefficient might be secured by expressing any of the other measures of deviation as a percentage of the average from which the deviations were computed. Pearson's coefficient has been generally adopted, however, and is the only one in wide use.

Measures of Skewness

Methods have been developed in the preceding sections for describing the central tendency of a frequency distribution and for measuring the degree of concentration or lack of concentration about that central tendency. One further measure is needed, and that is one which indicates the degree of skewness or asymmetry of a given distribution. For it is essential to know, in regard to a given distribution, whether the observations are arranged symmetrically about the central value, or are dispersed in an uneven, asymmetrical fashion about that value. Having such a figure it will be possible effectively to summarize the characteristics of a frequency distribution in three simple terms — an average, a measure of dispersion and a measure of skewness. Several such measures have been evolved.

If a frequency curve is perfectly symmetrical, mean, median and mode will coincide. As the distribution departs from symmetry these three values are pulled apart, the difference between the mean and the mode being greatest. This difference may be used, therefore, as a measure of skewness. It is desirable in this case, as in measuring relative variability, to secure an index in the form of an abstract number, which may be compared with similar figures derived from other distributions. To this end, Pearson has proposed dividing the absolute difference between mean and mode by the standard deviation of the given distribution. His formula is

$$sk \text{ (skewness)} = \frac{M - Mo}{\sigma}.$$

In a symmetrical distribution, where mean and mode coincide, the value of this measure will obviously be zero. The value may be positive or negative, depending upon the relative positions of the two averages on the scale.

For moderately skewed distributions the degree of skewness may be computed more readily from the formula

$$sk = \frac{3(M - Md)}{\sigma}.$$

This corresponds approximately to the other formula, because of the fact that in a moderately asymmetrical distribution the median lies between the mean and the mode, about one third of the distance from the former towards the latter.

Because it is difficult to locate the mode by simple methods, a measure of skewness more easily computed than Pearson's is desirable in some cases. Bowley has proposed such a method, based upon the relationship between the first and third quartiles and the median. If the distribution is symmetrical these two quartiles will be equidistant from the median; with an asymmetrical distribution this is not so. Therefore, if we let q_2 represent the difference between the upper quartile and the median and q_1 represent the difference between the median and the lower quartile, we may use the formula

$$sk = \frac{q_2 - q_1}{q_2 + q_1}$$

as a means of securing a measure of skewness. This value will vary between 0 and $\pm$ 1. For with perfect symmetry $q_2 = q_1$, and the measure is 0; with asymmetry so pronounced that the median and one of the quartiles coincide, either q_2 or q_1 becomes equal to zero, and the formula gives a value of + 1 or − 1. Bowley suggests that a value of .1 indicates a moderate degree of skewness, while a value of .3 indicates marked skewness.

The values secured from this measure are not, of course,

comparable with the values secured from the application of Pearson's formula for measuring skewness.

Kurtosis

Reference has been made to a fourth characteristic of a frequency curve which may be measured. This is the degree of flat-toppedness, as compared with the normal curve. A measure of *kurtosis*, the technical term for this characteristic, is given in Chapter XV.

REFERENCES

Bowley, A. L. *Elements of Statistics* (110–117).
Chaddock, R. E. *Principles and Methods of Statistics* (Chap. IX).
Davies, G. R. *Introduction to Economic Statistics* (29–46).
Jones, D. C. *A First Course in Statistics* (42–51).
Kelley, Truman L. *Statistical Method* (70–82).
King, W. I. *Elements of Statistical Method* (141–158).
Pearl, Raymond. *Medical Biometry and Statistics* (272–278).
Rietz, H. L. (editor). *Handbook of Mathematical Statistics* (27–33).
Rugg, H. O. *Statistical Methods Applied to Education* (149–179).
Secrist, Horace. *An Introduction to Statistical Methods* (377–424).
West, Carl J. *Introduction to Mathematical Statistics* (45–58).
Yule, G. U. *Introduction to the Theory of Statistics* (133–156).

CHAPTER VI

INDEX NUMBERS OF PRICES

The Nature of Index Numbers

The term "index number" has been applied to a number of somewhat similar devices employed in the analysis of statistical series. Index numbers have been most widely used in the study of price changes, but a brief consideration of certain other uses may make clear the essential characteristics of such measures. In its simplest form this name is applied to a term in a time series expressed as a relative. Thus an index number of cotton consumption in the United States might take the following form:

Table 33

Domestic Cotton Consumption in the United States, 1913–1923

(Consumption in 1913 = 100)

Year	*Cotton Consumption* (*Unit: Running Bale*)	*Cotton Consumption Relative*
1913	5,583,468	100.0
1914	5,448,760	97.6
1915	6,008,984	107.5
1916	6,620,415	118.6
1917	6,815,811	122.0
1918	6,176,547	110.5
1919	5,919,520	106.0
1920	5,843,200	104.5
1921	5,406,721	96.8
1922	6,087,520	109.0
1923	6,513,696	116.7

Similarly the price of a commodity may be expressed as a relative, the price at a given date or for a given period serving as base.

TABLE 34

Average Price of No. 1 Northern Spring Wheat, Minneapolis, 1913, 1919–1923

(Average Price in 1913 = 100)

Year	*Average Price per Bushel*	*Relative Price*
1913	$0.8735	100
1919	2.5660	294
1920	2.5581	293
1921	1.4660	168
1922	1.3450	154
1923	1.1810	135

The representation of the terms in a time series as relatives, with reference to a fixed base, permits a ready comparison of the values for different dates and permits the trend of the series to be perceived much more easily than when the data are presented in their original form. Comparison of the trends of different series is also facilitated.

Though the term index number has been applied to such relatives it is better practice to reserve the term for figures which represent the combination of a number of series. The series to be combined may relate to prices, production, consumption, wages, volume of trade or to any factor subject to temporal variation. Quite complex problems may be involved in the construction of any one of these special forms of index numbers, but the essential aim in all cases is to secure a single, simple figure which will indicate the net resultant of the forces acting on the constituent series.

A simple index number may be constructed to represent the course of coal and petroleum production in the United States. In the making of such an index it is necessary to combine in some way production figures for bituminous and anthracite coal and petroleum. The following are the production figures and the corresponding relatives for the three series, from 1910 to 1923:

TABLE 35

Production of Bituminous and Anthracite Coal and Petroleum in the United States, 1910–1923

(Production in 1910 = 100)

Year	*Prod. of Bit. Coal Million sh. tons*	*Rel.*	*Prod. of Anthr. Coal Million sh. tons*	*Rel.*	*Prod. of Petrol. Mill. bbls.*	*Rel.*
1910	417.1	100	84.49	100	209.6	100
1911	405.9	97	90.46	107	220.4	105
1912	450.1	108	84.36	100	222.9	106
1913	478.4	115	91.52	108	248.4	118
1914	422.7	101	90.82	107	265.8	127
1915	442.6	106	89.00	105	281.1	134
1916	502.5	120	87.58	103	300.8	143
1917	551.8	132	99.61	118	335.3	160
1918	579.4	139	98.83	117	355.9	170
1919	465.9	112	88.09	104	378.4	181
1920	568.7	136	89.60	106	442.9	211
1921	415.9	100	90.44	107	472.2	225
1922	404.5	97	52.90	63	557.5	266
1923	545.3	131	95.2	113	725.7	346

A rough index of fuel production, based upon these three series, is desired. It is impossible, obviously, to add the original figures, as the units are not the same. This difficulty may be avoided by using the relative figures, and a simple average of the three relatives for a given year may serve as the required index. The following is the series of index numbers thus secured:

TABLE 36

Index Numbers of Coal and Petroleum Production in the United States, 1910–1923

(Production in 1910 = 100)

Year	*Index*	*Year*	*Index*
1910	100	1917	137
1911	103	1918	142
1912	105	1919	132
1913	114	1920	151
1914	112	1921	144
1915	115	1922	142
1916	122	1923	196

In securing this index, by adding the three relative figures for a given year and dividing by three, equal weight has been given to each of the three series. Such an index of equally weighted relatives has been termed an unweighted index, but the term is misleading. Weights are used, the weights in this case being equal. It is clear that this index based upon equal weights does not reflect faithfully the three series combined in the present instance. For the three series are not of equal importance, as the system of equal weights assumes. The following figures showing the wholesale values in exchange in 1921 of bituminous coal, anthracite coal and petroleum indicate the relative importance of the three series: [1]

Mineral	*Wholesale Value in Exchange in* 1921
Bituminous Coal.....	$1,948,000,000
Anthracite Coal.....	731,000,000
Petroleum.....	712,000,000

Roughly, these stand to each other in the relation of 3, 1 and 1, and these weights may be assigned to the series under consideration. An index for each year may be computed, using these weights. The following example, showing the calculations for the years 1910 and 1923, will illustrate the method:

TABLE 37

Computation of Weighted Index Numbers of Coal and Petroleum Production

Mineral	*Relative Production,* 1910	*Wt.*	*Wt.* × *Rel.*	*Relative Production,* 1923	*Wt.*	*Wt.* × *Rel.*
Bituminous Coal..	100	3	300	131	3	393
Anthracite Coal...	100	1	100	113	1	113
Petroleum.....	100	1	100	346	1	346
		5	500		5	852

Index of fuel production, 1910 = 500 ÷ 5 = 100

Index of fuel production, 1923 = 852 ÷ 5 = 170

[1] The figures have been compiled by the U. S. Bureau of Labor Statistics.

The value of the index thus secured for each of the fourteen years covered is shown below:

TABLE 38

Weighted Index Numbers of Coal and Petroleum Production in the United States, 1910–1923

Year	*Index*	*Year*	*Index*
1910	100	1917	135
1911	101	1918	141
1912	106	1919	124
1913	114	1920	145
1914	107	1921	126
1915	111	1922	124
1916	121	1923	170

Quite important differences between the two series of index numbers are apparent. The second series, which is the more logically weighted, is, of course, the more accurate of the two, and gives a more faithful representation of the combined effect of the forces affecting the three series.

Another type of index number is one in which the items in the constituent series are totaled, the aggregate figure, instead of an average, serving as the representative of the entire group. Such a form of index number may be constructed only when the different series are all expressed in the same unit. This form is frequently employed as an indication of changes in the level of prices, the aggregate cost of a bill of goods at one period being compared with the aggregate cost of the same goods at other dates. The following figures illustrate this type of index.

TABLE 39

Bradstreet's Index Numbers of Wholesale Prices in the United States, 1913–1923

Year	*Index*	*Year*	*Index*
1913	$ 9.21	1919	18.66
1914	8.90	1920	18.81
1915	9.85	1921	11.37
1916	11.82	1922	12.12
1917	15.64	1923	13.40
1918	18.71		

Each of the yearly aggregates quoted above is the sum of the average prices during the year of 96 commodities at wholesale. Before being added all the prices are reduced to the "per pound" basis, so that a certain degree of comparability is secured. Such an index may be readily changed to the relative form, any year being taken as a base and the totals for the other years expressed as percentages of the figure for the base year.

The examples which have been given will indicate some of the many forms which index numbers may take. The term may refer to a simple relative number; it may be applied to an average of relative terms, or to an aggregate of relative or absolute figures. In all the examples given the index has been designed to serve as a measure of change over a period, as an indicator of changes in the values of time series. The term may have a much broader meaning than this. An index of the ability of salesmen might be constructed by giving numerical values to the factors determining their usefulness and securing an average of these values. An index of the efficiency of different departments in a business enterprise might be constructed. In any case, the construction of an index involves the reduction to comparable terms of a number of different factors and the replacement of these several terms by a single figure which may serve as their representative. Comparison is thus facilitated, whether it be comparison over time or comparison with indices secured by averaging terms relating to a similar unit. In all its forms (except the first limited and exceptional meaning in which it applies to a simple relative) an index number is thus a type of statistical average, and such numbers, in their construction and use, are subject to all the rules and limitations set forth in the development of the subject of averages.

In the present work we are interested only in the application of the index number device to time series. So varied, however, are the rules and practices relating to its applica-

tion to different types of time series that certain of these types must be treated separately. Our first concern is with index numbers of wholesale prices.

Price Changes

When price movements are surveyed in detail it is difficult to perceive order, or any definite trend. We find a multiplicity of conflicting movements. The following price quotations, taken at random, are roughly typical of what would be found were the entire field of prices canvassed in order to compare price movements from month to month:

Table 40

Commodity Prices at Wholesale [1]

Commodity	*Unit*	*Price (Wholesale) October, 1923*	*Price (Wholesale) November, 1923*
Brick, common building, average of yard prices	1000	$14.752	$14.746
Pig Iron, basic, Valley furnace	Gross ton	23.500	20.875
Cement, Portland, average of plant prices	Bbl.	1.893	1.842
Linseed Oil, raw, N. Y.	Gal.	.943	.910
Steel Billets, Bessemer, Pitts	Gross ton	40.000	40.000
Tin Plate, Pitts	100 lbs.	5.500	5.500
Copper, electrol., refinery	Pound	.126	.128
Lead, N. Y.	Pound	.069	.069
Zinc, N.Y.	Pound	.067	.067
Coal, Anthr. Chestnut, N.Y.	Gross ton	11.471	11.478
Coal, Bit. Mine run, Chi.	Net ton	4.600	4.525
Crude Petroleum, Penn., at wells	Bbl.	2.500	2.388
Gasolene, Motor, N. Y.	Gal.	.185	.170
Cotton, Middling, N. O.	Pound	.292	.339
Wheat, #2 Red Winter, Chi.	Bu.	1.097	1.061
Sugar, Granulated, N.Y.	Pound	.090	.087

Of the sixteen commodities listed, four showed no price change at all between October and November, 1923, three

[1] As compiled by the United States Bureau of Labor Statistics.

showed price increases, and in nine cases prices declined. Some of the price movements were inconsiderable, while some marked very material changes. Such, as seen here in miniature, is what happens in the price system as a whole. All prices do not, with absolute uniformity, move up or down or remain constant. Each of the thousands of commodities traded in on the markets of any country, or of the world, moves in its own individual way, subject to a variety of influences. Yet it does not act in isolation. In its price movements it affects other commodities, and is affected by them. And, in addition to the forces peculiar to each commodity, there are broad forces which act throughout the price system, affecting all commodities. It is the business of the economic statistician to bring order out of the chaos of price movements taking place at any given time and, out of the multiplicity of minor movements, to pick the broad trends which affect the whole economic system.

The forces which bring about the price movements which are to be studied are numerous and complicated, but some general conclusions may be drawn with regard to them. There are, in the first place, all those changes in production and consumption conditions peculiar to individual commodities and affecting directly the prices of those commodities. The opening of new fields, improvements in production technique in individual cases, changes in fashion and the transfer of demand from some commodities to others, changes in demand and supply with the seasons — all these are causing constant price readjustments. These are the changes which in ordinary times are most obvious, which are brought home directly to the individual merchant or consumer. Such changes affect the whole price system, as has been pointed out, but not in general by causing upward or downward movements in the system as a whole.

Such general movements are due to forces which are broader in their scope. The general improvement in production technique and the increase in the productivity of

human labor which has resulted have, by increasing the supply of commodities available for consumption, affected prices. Changes in monetary systems and, in particular, changes in the gold supply have exerted a direct and immediate influence upon prices, by affecting the supply of money in circulation. Similar in character have been changes in banking and credit systems and changes in commercial practice which have affected the use of credit instruments and the rapidity of circulation of money and credits. All these forces affect prices, though their incidence is not so specific as that of the factors affecting individual commodities directly.

Purpose of General Index Numbers of Wholesale Prices

These separate forces cannot be isolated and evaluated. Their joint action causes a perplexing variety of price changes. In studying these changes the problem might be approached from several different points of view. It might be desired to study the readjustments which take place within the price system, to determine the nature and degree of the shifts which changing conditions cause within the system. Such a study would yield valuable information as to the behavior of prices and the character of their interrelations. Our immediate problem, however, is the determination of the net resultant of all these forces. Do all price movements cancel each other so that while some prices move up and some down there is no net change? Or is there at a given time a preponderance of movements in one direction, causing the level of general prices to move upward or downward? If there is such a trend, what is it, and how may it be measured? Are the statistical methods which have been explained in the earlier sections applicable to the solution of this problem?

The first step in this study involves the answering of the

last question asked. It has been brought out that methods of summarizing quantitative data have been developed, but that these methods are applicable only when certain conditions are fulfilled. An average, it was noted, has no significance unless it represents a distinct central tendency in a mass of homogeneous data. Moreover, the type of average to be employed depends upon the character of the distribution it is to represent. Until the distribution of the original data is studied no average or other statistical measure can be intelligently employed. We must first, then, determine what the raw materials of the problem are, and study the frequency distribution secured when these raw materials are organized.

For the present a quite general purpose will be assumed, the determination of the change in the level of general wholesale prices between two specific dates. This is equivalent, of course, to measuring the change in the value of money between two given dates. The raw materials of the problem consist of a number of price quotations on individual commodities, quotations being secured for the two dates which are to be compared. Each pair of quotations measures the change in the price of a single commodity, a change caused by the interplay of many forces. When a great many such price quotations are brought together we have a mass of data representing the interaction of a multitude of forces, some individual and specific in their incidence, some general, affecting the prices of large groups of commodities or of all commodities. What we seek to determine is the net resultant of all these factors which are affecting prices. We seek a measure of the composite effect of the numerous forces which are causing individual prices to rise or fall. This measure will constitute an index number of wholesale prices.

The unit with which we must deal is a single price variation. Whether the statistical methods with which we are familiar may be employed in the organization and analysis

of a number of such units depends upon the behavior of such units in mass. The following examples illustrate the frequency distributions secured when these data are classified.

Frequency Distributions of Price Ratios

Each price variation is, of course, a ratio, the ratio of the price of a commodity at a given date to the price of the commodity at another date. The ratios may be reduced to a comparable basis by putting them all in the form of relatives, of the type illustrated in the earlier examples of index numbers. Thus, using one of the pairs of price quotations given above, the ratio of the price of pig iron in November, 1923, to the price in October, 1923, is \$20.875: \$23.500 which, in the form of a relative, becomes 88.8 : 100. In constructing the following frequency table, the prices at wholesale in 1914 of 346 commodities were expressed as relatives, with the 1913 price as a base in each case. The distribution of these 346 relative numbers is as follows:[1]

Table 41

Distribution of the Relative Prices of 346 Commodities in 1914

(Average prices in 1913 = 100)

Relative Prices	*Mid-points* m	*No. of cases* f	*Percentage of total number of cases*
62.5– 67.4	65	1	.3
67.5– 72.4	70	1	.3
72.5– 77.4	75	5	1.5
77.5– 82.4	80	7	2.0
82.5– 87.4	85	20	5.6
87.5– 92.4	90	35	10.0
92.5– 97.4	95	51	14.5
97.5–102.4	100	134	39.0
102.5–107.4	105	50	14.5

(*Continued on next page.*)

[1] The 346 commodities included were those employed by the U. S. Bureau of Labor Statistics in the construction of its index of wholesale prices. The original figures, and the relatives, appear in *Bulletin 269*, of that Bureau, on "Wholesale Prices, 1890–1919."

TABLE 41 (*Continued*)

Relative Prices	*Mid-points* *m*	*No. of cases* *f*	*Percentage of total number of cases*
107.5–112.4	110	21	6.0
112.5–117.4	115	12	3.5
117.5–122.4	120	3	1.0
122.5–127.4	125	2	.6
127.5–132.4	130	2	.6
132.5–137.4	135		
137.5–142.4	140	1	.3
142.5–147.4	145		
147.5–152.4	150	1	.3
		346	100.0

The frequency polygon representing this distribution appears in Fig. 47. For purposes of comparison with similar

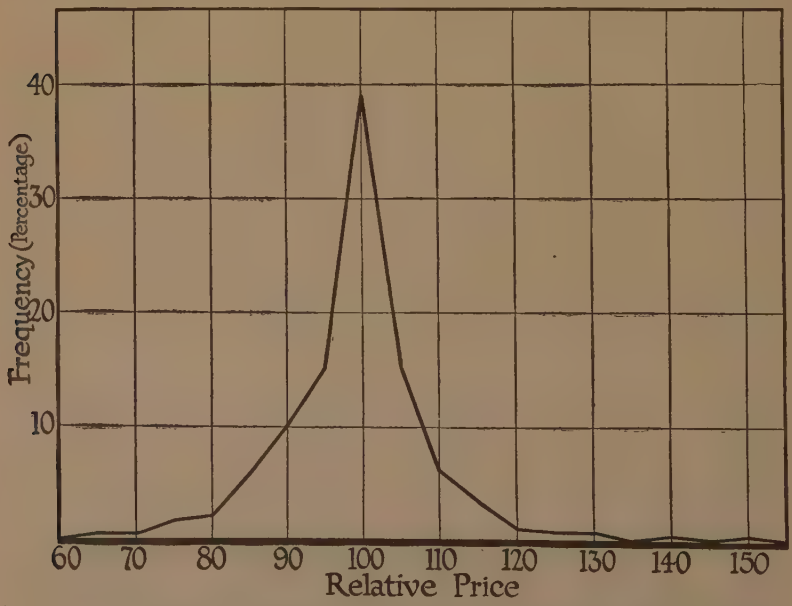


FIG. 47. — Frequency Polygon: Distribution of Relative Prices of 346 Commodities in 1914. (Average prices in 1913 = 100)

distributions the figure shows the percentage distribution. The correspondence of this frequency distribution to the standard types portrayed in earlier sections is obvious.

There is the same marked concentration about a central tendency, in this case a tendency of prices to remain stable, for 39% of all the cases showed a change not exceeding 2.5% from their prices in the base year. There is also, in this case, a fairly symmetrical distribution about this central tendency, though the range above the mode is greater than the range below, a fact of considerable significance. Without at present considering the question as to which average might best be used to represent the central tendency in this distribution, it is apparent that the use of some average is quite legitimate.

The example just given has been based upon price variations from one year to the next, over a period during which the level of general prices showed no change. W. C. Mitchell gives a much more comprehensive illustration, based upon the distribution of 5578 price variations from one year to the next over the period 1890–1913, which shows the same general grouping. The excess of the range above the mode over the range below is somewhat more pronounced, in connection with which fact it should be noted that prices were rising during most of the 23 years covered. The distribution secured by Mitchell is shown in Fig. 41.

The inertia of prices is most conspicuous when year to year price changes are studied. It is therefore advisable to consider the character of price variations over a longer period, that we may learn whether the same type of distribution is secured. Two examples are given, one of price changes over a ten-year period, a period so chosen that the price level at the end was approximately the same as at the beginning; the other of price changes over a five-year period characterized by rapidly rising prices. The table following shows the distribution of 222 price variations, prices in 1900 being expressed as relatives on an 1890 base. The general price level, it should be noted, fell from 1890 to 1896 and rose from 1896 to 1900; prices in 1900 were approximately one per cent lower than in 1890.

TABLE 42

Distribution of Relative Prices of 222 Commodities in 1900

(Average prices in 1890 = 100)

Relative Prices	*Mid-points* m	*No. of cases* f	*Percentage of total number of cases*
42.5– 47.4	45	2	1.0
47.5– 52.4	50	1	.5
52.5– 57.4	55	3	1.5
57.5– 62.4	60	3	1.5
62.5– 67.4	65	3	1.5
67.5– 72.4	70	3	1.5
72.5– 77.4	75	8	3.5
77.5– 82.4	80	12	5.0
82.5– 87.4	85	19	8.5
87.5– 92.4	90	20	9.0
92.5– 97.4	95	30	13.5
97.5–102.4	100	27	12.0
102.5–107.4	105	24	11.0
107.5–112.4	110	14	6.0
112.5–117.4	115	18	8.0
117.5–122.4	120	10	4.5
122.5–127.4	125	7	3.0
127.5–132.4	130	9	4.0
132.5–137.4	135	2	1.0
137.5–142.4	140	4	2.0
142.5–147.4	145	1	.5
147.5–152.4	150	1	.5
152.5–157.4	155	1	.5
		222	100.0

These data are plotted in the form of a frequency polygon in Fig. 48, the percentage distribution being shown. (In comparing Figures 47 and 48 the difference in the vertical scales should be noted.)

The distributions depicted in Figs. 47 and 48 differ materially. The range of the variations is greater in the second case, a condition naturally to be expected because of the longer period covered. Secondly, a very much smaller percentage of cases is concentrated in the modal group, though there is still a pronounced central tendency. Both

distributions, as plotted on the arithmetic scale, are fairly symmetrical. (In considering this fact, it should be remembered that the price level as measured by the arithmetic mean was practically the same at the two dates compared in each case.) In the first case the concentration about the central tendency is much more marked, and the deviations of individual price ratios from the arithmetic

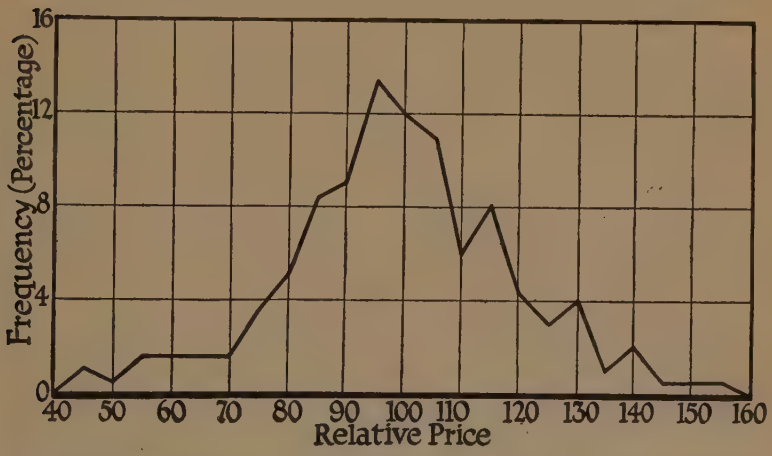


FIG. 48. — Frequency Polygon: Distribution of Relative Prices of 222 Commodities in 1900. (Average prices in 1890 = 100)

mean are smaller. This distribution resembles one which would be secured from highly accurate physical measurements, or the distribution of shots from a very accurate piece of artillery. The second curve corresponds to one representing less accurate physical measurements, or to the distribution of shots from an old or inaccurate field piece. The modal value occurs less frequently and the deviations from the mean are greater. It has been established that the longer the period covered in price comparisons such as those made above, the more pronounced is the tendency shown in the second curve. The value of the maximum ordinate falls and the range of the distribu-

tion increases. The curve becomes flatter and more extended as the time interval increases. And, quite obviously, as this process goes on the representative character of any type of average declines. Unless there is concentration about a central tendency an average is merely an abstraction, without concrete significance.

It is possible at this point to state as a tentative conclusion that price variations are capable of statistical measurement, that they may be appropriately represented by an average value, provided the period covered is not too long. No definite statement can be made as to the maximum period over which price variations may be measured. Index numbers having accurate and significant values must be based upon comparisons over relatively short periods, the most accurate being year-to-year comparisons. Index numbers designed merely to show general trends in prices may cover longer periods, though the makers and users of such index numbers should realize their limitations.[1]

The two examples given above represented comparisons between dates when prices, as measured by the arithmetic mean of a number of variations, were at approximately the same level. These are, of course, particular cases, and it is desirable to study the distribution of price variations over a period characterized by changes in the general level of prices. The following table shows the distribution of the relative prices of 1437 commodities in 1918, average prices during the period July, 1913, to June, 1914, serving as base.[2]

[1] These conclusions are largely based upon studies made by W. C. Mitchell. Cf. "The Making and Using of Index Numbers," *Bulletin 284* (Wholesale Price Series), United States Bureau of Labor Statistics.

[2] Computed by the Price Section of the War Industries Board; reproduced in Part 1, *Bulletin 284*, U. S. Bureau of Labor Statistics, 70.

TABLE 43

Distribution of the Relative Prices of 1437 Commodities in 1918

(Average Prices, July, 1913, to June, 1914 = 100)

Relative Prices	*Mid-points* m	*No. of cases* f	*Percentage of total number of cases*
36	36	1	*
49	49	1	*
50– 69	60	4	.3
70– 89	80	17	1.2
90–109	100	61	4.3
110–129	120	64	4.5
130–149	140	130	9.0
150–169	160	212	14.7
170–189	180	219	15.2
190–209	200	164	11.4
210–229	220	135	9.4
230–249	240	104	7.2
250–269	260	76	5.3
270–289	280	54	3.8
290–309	300	42	3.0
310–329	320	30	2.1
330–349	340	31	2.1
350–369	360	16	1.1
370–389	380	13	.9
390–409	400	7	.5
410–429	420	7	.5
430–449	440	8	.6
450–469	460	4	.3
470–489	480	4	.3
490–509	500	4	.3
510–529	520	5	.4
530–549	540	3	.3
550–569	560	4	.3
587	587	1	*
627	627	1	*
727	727	1	*
730	730	1	*
743	743	1	*
761	761	1	*
784	784	1	*
826	826	1	*
848	848	1	*
900	900	1	*
1165	1165	1	*

(*Continued on next page.*)

TABLE 43 (*Continued*)

Relative Prices	*Mid-points* m	*No. of cases* f	*Percentage of total number of cases*
1356	1356	1	*
1585	1585	1	*
1764	1764	1	*
2049	2049	1	*
2863	2863	1	*
3009	3009	1	*
		1437	

* Less than one tenth of one per cent.

This distribution is shown graphically in Fig. 49. (The scales in this figure are not the same as those employed in the two figures preceding).

A study of this distribution bears out the conclusion

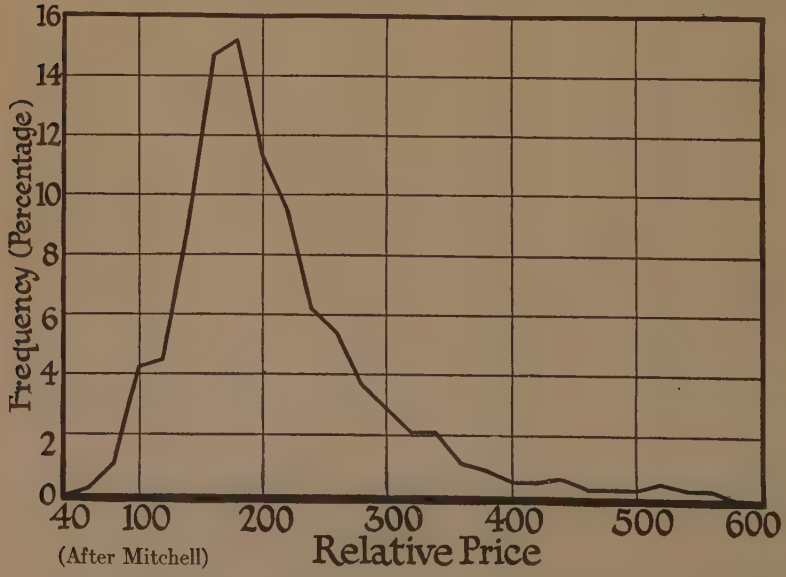


FIG. 49. — Frequency Polygon: Distribution of Relative Prices of 1437 Commodities in 1918. (Average prices, July, 1913, to June, 1914 = 100)

reached from the two examples preceding. There is a central tendency sufficiently pronounced to be well represented by an average. In this case, moreover, the modal

group is that with a mid-point of 180, so that the tendency toward concentration cannot be attributed to inertia, but to the presence of external forces affecting the price system as a whole. There is, however, one marked point of difference between this distribution and the two others. The tendency toward skewness, which was in evidence in the first example, is pronounced in this case. The curve, as plotted on the arithmetic scale, is markedly asymmetrical. The greatest concentration is near the lower limit of the scale and a long tail, extending in fact far beyond the limit of the chart, tapers out to the right. The details of this condition are most clearly apparent from the table. Starting from 100 as a base, the smallest relative price is 36, an absolute decline of 64 points on the scale. The highest relative price is 3009, an absolute increase of 2909 points. It is true that this occurred during a period characterized by a pronounced rise in the general price level. (The index of the War Industries Board shows that prices in 1918 were 94% above those in the base period.) That accounts for the location of the modal group, but it does not explain the nature of the distribution. In explaining the shape of the frequency curve we shall be preparing to answer a question which is fundamental in index number construction, namely, What type of average should be employed in securing a figure representative of the central tendency in a distribution of price variations?

The Problem of Averaging Price Variations

A price increase, expressed as a relative, has no upper limit. An increase of 100, 500, 1000 per cent or more is conceivable and possible. The greatest price increase noted by the War Industries Board in its study of prices during the war was one of 4981 per cent, in the case of acetiphenetidin. But 100 per cent is the maximum decline possible, as that would mean that the price of a commodity had fallen

to zero. This is the explanation of the skewness noted in the curves shown. When any considerable number of price ratios are tabulated the corresponding frequency curve, plotted on an arithmetic scale, shows this characteristic feature, a feature which is most conspicuous during a period of rising prices.

The bearing of this fact on the selection of an appropriate type of average has been brought out in the discussion of averages. When ratios are being averaged the geometric mean is the appropriate average to employ. This fact is clearly demonstrated by comparing the different averages for the distribution last shown. The arithmetic mean of the 1437 price ratios there tabulated is 217, while the mode is 180. When a mean departs any considerable distance from the mode it obviously loses something of its representative character. The geometric mean in this case is 194, a value much closer to both the mode and the median (191). Quite apart from other arguments, this relationship between the different averages indicates that the geometric mean is to be preferred to the arithmetic, in the present case.

The fact was noted in the chapter on averages that, when a frequency distribution plotted on a logarithmic scale approximates the normal curve more closely than when plotted on the arithmetic scale, this indicates that the data follow the geometric rather than the arithmetic law of dispersion. Figure 50 shows the distribution of the 1437 relative prices in 1918 on the 1913–14 base, when the relative prices are plotted on a logarithmic scale. The pronounced asymmetry which is apparent when the data are plotted on an arithmetic scale is gone, and a curve very closely resembling the normal type is secured.

While this tendency of price ratios to follow a geometric law of dispersion is a strong argument for the use of the geometric mean in averaging these ratios, this does not mean that price index numbers may not be constructed by other

methods. In general, the geometric mean is the logical average to use, but good index numbers may be constructed in other ways. This point will be developed in a subsequent section.

The argument which has been developed in the preceding pages may be briefly summarized. Before discussing the

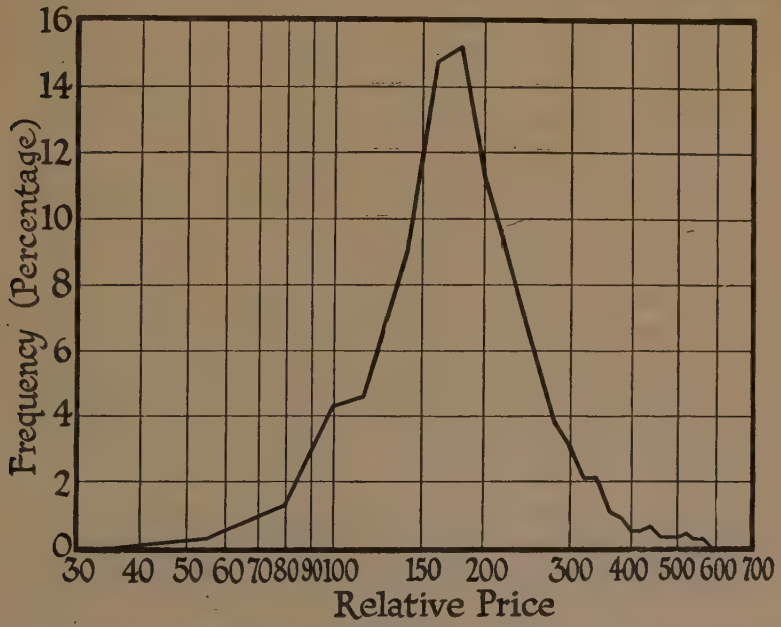


FIG. 50. — Frequency Polygon: Distribution of Relative Prices of 1437 Commodities in 1918, with Relative Prices plotted on Logarithmic Scale. (Average prices, July, 1913, to June, 1914 = 100)

practice of index number construction it was considered advisable to study the character of the raw materials and the nature of the distributions secured when these materials are brought together, in order to determine whether ordinary statistical methods are appropriate. The raw materials, we have seen, consist of individual price variations, expressed as ratios. When a number of these ratios are assembled a frequency distribution is secured which

somewhat resembles the distribution of data following the normal law of error. A central tendency, which may legitimately be represented by an average, is apparent in the distribution of price variations. The central tendency is less marked, however, and the deviations from it are more pronounced, the longer the period covered in the price comparison, so that an average becomes less representative as this period increases. In addition, a tendency toward skewness has been noted, and this was seen to be quite pronounced in a period of rising prices. This skewness is due to the fact that we are dealing with ratios which have a definite lower limit and no upper limit. Referring to the previous discussion of averages, we have found that price variations constitute a type of material following the geometric law of dispersion, and that if we are seeking the average of these variations the geometric mean should be employed. While this general conclusion holds true, it has been noted that for certain purposes other methods of index number construction may appropriately be employed.

The problem as to the type of average which should be used, and the connected problem as to the best method of weighting, have occupied the center of the stage in all discussions on the theory of index number construction. Assuming that the purpose of an index number of wholesale prices is to serve as a measure of the average ratio of change in general prices, the best theoretical solution of the first problem has been suggested. The question of weighting introduces a new element.

The chief reason for the employment of a system of weights in the construction of index numbers has been indicated above. Whether the index number is to measure variations in the physical volume of production, in the volume of trade, or in prices, all the quotations and ratios which must be averaged in securing the index number are not of equal importance. In the construction of an index of production coal is more important than lead; in the con-

struction of an index of wholesale prices the price of wheat is more important than the price of flaxseed. The problem of measuring the relative importance of the different commodities to be included in an index number is one which has been solved in various ways. In large part the selection of weights is dependent upon the particular purpose which a given index number is to serve, as well as upon practical difficulties involved in securing the data upon which weights may be based. Several examples of varying weights are shown in the section which follows. A further discussion of the subject will follow the consideration of these examples.

Variety of Methods Employed In Index Number Construction

Many methods have been and are being employed in the construction of index numbers of wholesale prices. Usage varies for many reasons. There are some differences of opinion as to which is theoretically the best method. There are practical difficulties to be surmounted, difficulties which inevitably cause differences in practice because of the varying resources of the agencies engaged in these tasks. And there are, finally, certain differences due to the varying purposes for which index numbers are constructed.

Prevailing differences in practice and differences in the results secured by the employment of various methods in the construction of index numbers can perhaps be illustrated most effectively by the application of a number of methods to the same data. The table which follows presents the raw material to which these various methods are to be applied — the average farm prices, on December 1, of twelve leading crops, from 1910 to 1923.

TABLE 44

Average Farm Prices, on December 1, of Twelve Leading Crops, 1910–1923 [1]

Crop	*Unit*	1910	1911	1912	1913	1914	1915	1916	1917	1918	1919	1920	1921	1922	1923
Corn	Bu.	$ .480	$ .618	$ .487	$ .691	$ .644	$ .575	$ .889	$ 1.279	$ 1.365	$ 1.347	$ .677	$ .423	$ .657	$ .727
Cotton	Lb.	.141	.088	.119	.122	.068	.113	.196	.277	.276	.356	.140	.162	.238	.310
Hay	Ton (short)	12.14	14.29	11.79	12.43	11.12	10.63	11.22	17.09	20.13	19.55	16.72	12.11	12.59	14.07
Wheat	Bu.	.883	.874	.760	.799	.986	.919	1.603	2.008	2.042	2.151	1.443	.926	1.009	.923
Oats	Bu.	.344	.450	.319	.392	.438	.361	.524	.666	.709	.715	.472	.302	.394	.415
Potatoes (white)	Bu.	.557	.799	.505	.687	.487	.617	1.461	1.228	1.193	1.606	1.164	1.101	.582	.823
Sugar [2]	Lb.	.0393	.0494	.0405	.0354	.0392	.0520	.0569	.0672	.0728	.0728	.0577	.0370	.0570	.0747
Barley	Bu.	.578	.869	.505	.537	.543	.516	.881	1.137	.917	1.210	.707	.419	.525	.540
Tobacco	Lb.	.093	.094	.108	.128	.098	.091	.147	.240	.280	.390	.211	.199	.231	.203
Flaxseed	Bu.	2.317	1.821	1.147	1.199	1.260	1.740	2.486	2.966	3.401	4.383	1.766	1.451	2.114	2.108
Rye	Bu.	.715	.832	.663	.634	.865	.834	1.221	1.660	1.516	1.345	1.278	.697	.692	.647
Rice	Bu.	.678	.797	.935	.858	.924	.906	.889	1.896	1.918	2.668	1.189	.952	.934	1.103

[1] For the period 1910–20 these figures were assembled by Warren M. Persons; cf. "Fisher's Formula for Index Numbers," *Review of Economic Statistics*, Prel. Vol. III, 105. For the period 1921–23 the prices quoted are from *Weather, Crops and Markets*.

[2] The price of sugar given for each year is the wholesale price of the raw product in the month of December. No figure corresponding to the farm price could be secured for this commodity.

Explanation of Symbols

The symbols to be employed in the computation of different types of index numbers have the following meanings:

p_0' : price of a given commodity at time "0" (the base period).
q_0' : quantity of same commodity at time "0".
p_1' : price of same commodity at time "1".
q_1' : quantity of same commodity at time "1".
p_0'': price of a second commodity at time "0".
q_0'': quantity of second commodity at time "0".
p_1'': price of second commodity at time "1".
q_1'': quantity of second commodity at time "1".
$\frac{p_1'}{p_0'}$: a price relative (relation of price of a given commodity at time "1" to price of same commodity at time "0").
$\frac{q_1'}{q_0'}$: a quantity relative.
P_0 : price level at time "0".
P_1 : price level at time "1".

Simple Index Numbers of Prices

In his exhaustive analysis of methods of index number construction [1] Irving Fisher distinguishes six fundamental types: the aggregative (or price aggregate), the arithmetic, harmonic, geometric, median and mode. The latter has never been employed in a practical way, and may be omitted. The characteristics of the five remaining types may be brought out by considering each of them in its simplest form, before examining the more complicated combinations.

Aggregates of Actual Prices

In the construction of index numbers of the simple aggregative type, commodity prices pertaining to a given

[1] *The Making of Index Numbers*, Houghton Mifflin Co., 1922.

date are added; general price changes are measured by comparing the results thus secured for different dates. Using the above symbols

$$\frac{P_1}{P} = \frac{\Sigma p_1}{\Sigma p_0}$$

When such index numbers are constructed from the data of Table 44 the following results are secured. The actual aggregates are given in column (2); to facilitate comparison the same figures are reduced to relatives, with the 1910 aggregate as base, in column (3).

TABLE 45

Index Numbers of Farm Crop Prices

(Aggregates of Actual Prices)

(1) *Year*	(2) *Index (aggregate of actual prices)*	(3) *Index, Relative (1910 = 100)*
1910.....	$18.9653	100
1911.....	21.5814	114
1912.....	17.3785	92
1913.....	18.5124	98
1914.....	17.4722	92
1915.....	17.3540	91
1916.....	21.5739	113
1917.....	30.5142	161
1918.....	33.8198	178
1919.....	35.7938	189
1920.....	25.8247	136
1921.....	18.7790	99
1922.....	20.0230	106
1923.....	21.9437	116

The results secured by this method of constructing index numbers of prices will be compared shortly with results secured from the same data by other methods. The chief weakness of this type of index number is obvious. This is not an unweighted nor yet an equally weighted index. The influence of each commodity upon the result is dependent upon the price of the unit in which it happens

to be traded. In the present index, hay, which is quoted by the ton, is given more weight than all the other 11 commodities combined, with flaxseed second in importance. The index secured by adding the quotations is weighted in an entirely illogical fashion and cannot be accepted as reflecting the course of farm crop prices.

One method which has been employed for avoiding the unequal weighting caused by the difference in units in which different commodities are traded is to reduce all quotations to the same unit. Thus hay, rice, corn, cotton and the other commodities might all be quoted by the pound, and these quotations added to secure the index. Yet this method, which has been employed in the construction of Bradstreet's index, merely replaces one system of illogical weighting by an equally illogical one. Equal weight, if such is desired, is not given to all commodities by this method. Thus, in 1910 hay was worth $.00607 per pound, cotton $.141 per pound and rice $.015 per pound, cotton having a weight in an aggregate of per pound prices 9 times that of rice and 23 times that of hay.

Arithmetic Averages of Relative Prices

Another method employed in the construction of index numbers involves the reduction of each quoted price to a relative, with reference to the price of the same commodity at a certain basic date, these relative figures then being averaged by any of the conventional methods. The following example illustrates the first phase of this process, data for two years being utilized. The year 1910 is taken as base.

TABLE 46

Computation of Relative Prices for the Construction of Index Numbers

(1) *Commodity*	(2) *Unit*	(3) *Price, 1910*	(4) *Relative*	(5) *Price, 1911*	(6) *Relative*
Corn	Bu.	$ .480	100	$ .618	128.8
Cotton	Lb.	.141	100	.088	62.4
Hay	Ton (short)	12.14	100	14.29	117.7
Wheat	Bu.	.883	100	.874	99.0
Oats	Bu.	.344	100	.450	130.9
White Potatoes	Bu.	.557	100	.799	143.5
Sugar	Lb.	.0393	100	.0494	125.6
Barley	Bu.	.578	100	.869	150.2
Tobacco	Lb.	.093	100	.094	101.1
Flaxseed	Bu.	2.317	100	1.821	78.7
Rye	Bu.	.715	100	.832	116.2
Rice	Bu.	.678	100	.797	117.5
			1200		1371.6

From these figures the arithmetic averages of relative prices in these two years may be readily computed. The formula for any single relative is $\frac{p_1'}{p_0'}$. When there are N relatives the formula for the index number at time "1" is

$$\frac{\Sigma\left(\frac{p_1}{p_0}\right)}{N}.$$

In the present case

$$\text{Index (1910)} = \frac{1200}{12} = 100.$$

$$\text{Index (1911)} = \frac{1371.6}{12} = 114.3.$$

Index numbers computed in this way for the years 1910 to 1923, inclusive, are shown in column (3) of Table 49.

WEIGHTS IMPLICITLY EMPLOYED IN SIMPLE AVERAGES OF RELATIVE PRICES

This type of index number is usually termed an "unweighted" index of relative prices. It is weighted, however,

just as are the types illustrated in the two examples preceding. The quantity employed as weight in each case is the amount of each commodity which would sell for $100 in the base year. In the preceding example the following quantities have been employed as weights:

Corn	208.3	bu.
Cotton	710.0	lbs.
Hay	8.24	tons
Wheat	113.3	bu.
Oats	291.0	bu.
Potatoes	180.0	bu.
Sugar	2650.0	lbs.
Barley	173.2	bu.
Tobacco	1076.0	lbs.
Flaxseed	43.2	bu.
Rye	140.0	bu.
Rice	147.7	bu.

What has been done, in effect, in the computation of the simple average of relative prices has been to determine the aggregate amount for which the above quantities would sell in each of the eleven years included. At 1910 prices each of the above quantities would sell for $100, the aggregate value being $1200; at 1911 prices the aggregate value of the above quantities was $1371.60. These aggregates, divided by 12, give the index numbers shown in column (3), Table 49: 100 for 1910, 114 (114.3) for 1911, etc. Thus the "unweighted average of relative prices" is in fact a weighted aggregate of actual prices. It is equally weighted in the sense that the value of the quantity of each commodity employed as weight was equal to $100 in the base year, 1910.[1]

Medians of Relative Prices

The median rather than the arithmetic mean may be employed in securing the average of the relative prices for

[1] Attention was called to this characteristic of the simple average of relative prices by F. R. Macaulay, *American Economic Review*, Dec., 1915, 928.

each year. When the relatives in column (6) of Table 46 are arranged in order of magnitude the following distribution is secured:

62.4	117.7
78.7	125.6
99.0	128.8
101.1	130.9
116.2	143.5
117.5	150.2

The smallest relative price is 62.4, the greatest 150.2; the median value is 117.6. This median value is the index number for 1911. All the index numbers computed in this way from the medians of relative prices are presented in column (4), Table 49.

Geometric Averages of Relative Prices

The geometric averages of the relative prices for the various years may now be computed and the results compared with those secured in the preceding examples. A single relative being represented by the symbol $\frac{p_1'}{p_0'}$, the formula for the geometric mean of N relatives is

$$M_g = \sqrt[N]{\frac{p_1'}{p_0'} \times \frac{p_1''}{p_0''} \times \frac{p_1'''}{p_0'''} \times \ldots}$$

A geometric mean is generally computed by the aid of logarithms; in this case

$$\text{Log } M_g = \frac{\log\left(\frac{p_1'}{p_0'}\right) + \log\left(\frac{p_1''}{p_0''}\right) + \log\left(\frac{p_1'''}{p_0'''}\right) + \ldots}{N}$$

The method of computation may be illustrated for the years 1910 and 1911. The relative prices of the various commodities are repeated from Table 46.

TABLE 47

Computation of Geometric Averages of Relative Prices

(1) *Commodity*	(2) *Relative Price, 1910*	(3) *Logarithm of Fig. in col. (2)*	(4) *Relative Price, 1911*	(5) *Logarithm of Fig. in col. (4)*
Corn	100	2.0	128.8	2.10992
Cotton	100	2.0	62.4	1.79518
Hay	100	2.0	117.7	2.07078
Wheat	100	2.0	99.0	1.99564
Oats	100	2.0	130.9	2.11694
White Potatoes	100	2.0	143.5	2.15685
Sugar	100	2.0	125.6	2.09899
Barley	100	2.0	150.2	2.17667
Tobacco	100	2.0	101.1	2.00475
Flaxseed	100	2.0	78.7	1.89597
Rye	100	2.0	116.2	2.06521
Rice	100	2.0	117.5	2.07004
		24.0		24.55694

$$\text{Log } M_g \text{ (1910)} = \frac{24}{12} = 2$$

M_g = antilogarithm of 2 = 100

$$\text{Log } M_g \text{ (1911)} = \frac{24.55694}{12} = 2.04641$$

M_g = antilogarithm of 2.04641 = 111.3.

This value, 111.3, is the index number for 1911. The results for all the years are summarized in column (5), Table 49.

HARMONIC AVERAGES OF RELATIVE PRICES

The characteristics of the harmonic average have been discussed in a preceding chapter. The reciprocal of the harmonic mean, it will be recalled, is the arithmetic mean of the reciprocals of the constituent measures. The constituent items, in the present case, are price relatives of the form $\frac{p_1'}{p_0'}$. The reciprocal of such a relative is $\frac{p_0'}{p_1'}$. The

formula for the harmonic mean of N price relatives is, therefore,

$$\frac{1}{H} = \frac{\frac{p_0'}{p_1'} + \frac{p_0''}{p_1''} + \frac{p_0'''}{p_1'''} + \cdots}{N}$$

or

$$H = \frac{N}{\Sigma\left(\frac{p_0}{p_1}\right)}.$$

The method of computation is illustrated in the following table:

TABLE 48

Computation of Harmonic Averages of Relative Prices

(1) *Commodity*	(2) *Relative price,* 1910	(3) *Reciprocal of Fig. in col.* (2)	(4) *Relative price,* 1911	(5) *Reciprocal of Fig. in col.* (4)
Corn	100	.01	128.8	.007763975
Cotton	100	.01	62.4	.01602564
Hay	100	.01	117.7	.008496177
Wheat	100	.01	99.0	.010101010
Oats	100	.01	130.9	.007639419
White Potatoes	100	.01	143.5	.006968641
Sugar	100	.01	125.6	.007961783
Barley	100	.01	150.2	.006657790
Tobacco	100	.01	101.1	.009891197
Flaxseed	100	.01	78.7	.01270648
Rye	100	.01	116.2	.008605852
Rice	100	.01	117.5	.008510638
		.12		.111328602

$$H\ (1910) = \frac{12}{.12} = 100$$

$$H\ (1911) = \frac{12}{.111328602} = 107.8.$$

The index numbers computed in this way for all the years included in the study are shown in column (6), Table 49.

In the construction of the five types of index numbers explained above no attempt has been made to use a logical weighting system. All are termed "unweighted" averages, a term which is quite misleading. The first index constructed, based on aggregates of actual prices, is a heavily weighted index number, though the weights are illogical. In the next four the quantities employed as weights are the amounts purchasable for $100 in 1910. The five results may now be brought together and compared. In each case the index is given to the nearest whole number.

TABLE 49

Index Numbers of Farm Crop Prices, 1910–1923

(1910 = 100)

(1) *Year*	(2) *Aggregates of actual prices (as relatives)*	(3) *Arithmetic averages of relative prices*	(4) *Medians of relative prices*	(5) *Geometric averages of relative prices*	(6) *Harmonic averages of relative prices*
1910	100	100	100	100	100
1911	114	114	118	111	108
1912	92	95	93	92	90
1913	98	104	98	100	97
1914	92	101	102	97	91
1915	91	104	104	102	101
1916	113	156	152	151	147
1917	161	208	208	204	198
1918	178	215	209	210	205
1919	189	252	226	241	231
1920	136	151	143	145	139
1921	99	115	99	107	101
1922	106	129	114	124	119
1923	116	142	134	135	129

These index numbers are plotted in Fig. 51.

COMPARISON OF SIMPLE INDEX NUMBERS

The four averages of relative prices agree much more closely with each other than with the index numbers based

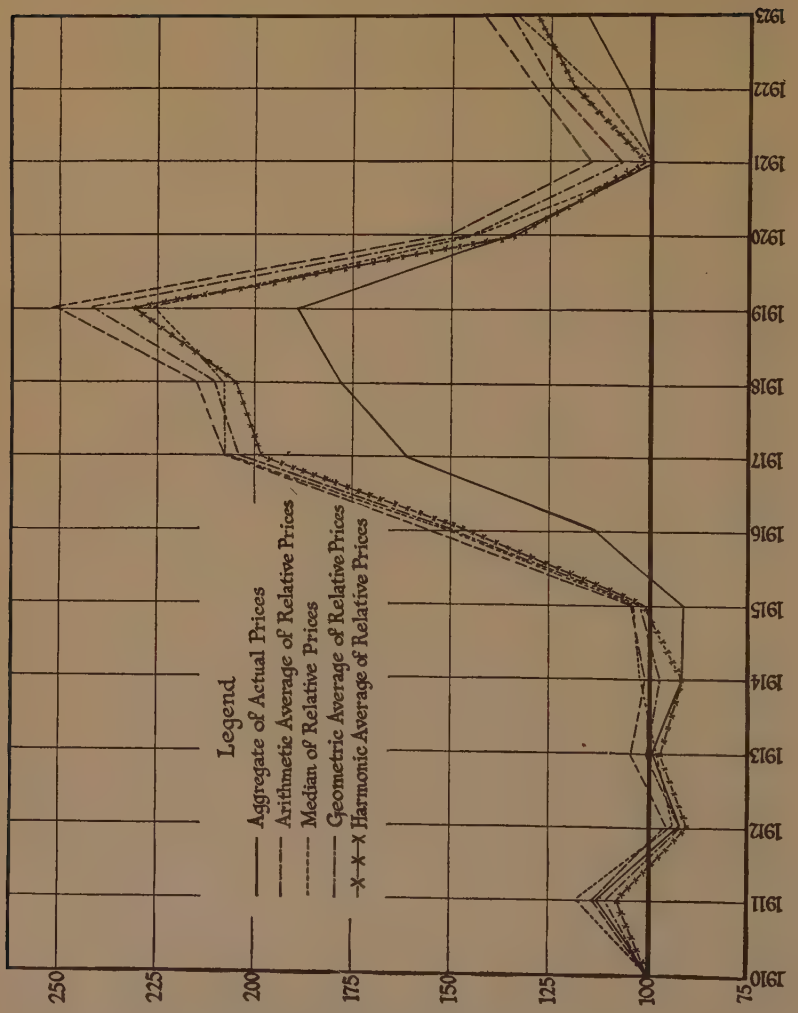


FIG. 51. — Comparison of Five Simple Index Numbers of Farm Crop Prices, 1910–1923. (1910 = 100)

on aggregates. For reasons already suggested the latter is quite untrustworthy as a measure of price changes. Of the other index numbers, the arithmetic, geometric and

harmonic means show a consistent relationship, a fact which follows from the nature of the averages employed. Except in the base year the geometric mean is always less than the arithmetic and the harmonic is always less than the geometric, the amount of difference increasing as the dispersion of prices becomes greater. The median, with only twelve items to be averaged, is somewhat unstable, and its relationship to the other averages is not always a consistent one.

How are we to choose among these varying results? No one of these "unweighted" index numbers is perfect, for weights which have crept in do not measure the relative importance of the various commodities included in the index numbers. But, neglecting for the moment the question of weights, is it possible to test the adequacy of the different methods for measuring changes in the prices as given?

The Time Reversal Test

For this purpose Irving Fisher has employed what he terms the "time reversal test." This is merely a test to determine whether a given method will work both ways in time, forward and backward. If from 1910 to 1911 sugar should increase from four to eight cents a pound, the price in 1911 would be 200 per cent of the price in 1910, and the price in 1910 would be 50 per cent of the price in 1911. One figure is the reciprocal of the other; their product ($2.00 \times .50$) is unity. Similarly, if a given method of index number construction shows the general price level in one year to be 200 per cent of the level in the preceding year, it should work correctly when reversed; it should show that the price level in the first year was 50 per cent of the price level in the second year. When the data for any two years are treated by the same method, but with the bases reversed, the two index numbers secured should be reciprocals of each

other. Their product should always be unity. If it is not, there is an inherent bias in the method.

This test may be applied to the methods employed above, using prices for 1910 and 1911. With 1910 as base the following results were obtained:

Year	*Aggregates of actual prices (as relatives)*	*Arithmetic averages of relative prices*	*Medians of relative prices*	*Geometric averages of relative prices*	*Harmonic averages of relative prices*
1910	100	100	100	100	100
1911	113.79414	114.3	117.6308	111.3	107.8

and with 1911 as base:

Year	*Aggregates of actual prices (as relatives)*	*Arithmetic averages of relative prices*	*Medians of relative prices*	*Geometric averages of relative prices*	*Harmonic averages of relative prices*
1910	87.87799	92.76	85.0117	89.85	87.47
1911	100	100	100	100	100

When the index numbers for 1911 in the first table are multiplied by the corresponding index numbers for 1910 in the second table, we have the following values. (In securing these products the index numbers are put in the ratio, not in the percentage form.)

Aggregates of actual prices	*Arithmetic averages of relative prices*	*Medians of relative prices*	*Geometric averages of relative prices*	*Harmonic averages of relative prices*
1.00	1.0602	1.00	1.00	.9429

This time reversal test is met by three of the methods employed. It is not met by either the arithmetic or har-

monic averages. The former has a distinct upward bias, amounting to more than six per cent when the errors for 1910 and 1911 are compounded, while the harmonic mean shows almost as large an error in the opposite direction. Unless the inherent bias which is found in both these averages is rectified in some way, methods based upon these averages should not be used in the construction of index numbers.

The Weighting of Index Numbers

Five simple index numbers of prices have been described in the preceding section. With the introduction of weighting the number of possible combinations is greatly increased, but only a few of these types need concern us here.

In the construction of an accurate measure of price changes logical weights must be employed, weights which truly reflect the relative importance of the commodities included. If the weighting problem is ignored haphazard and illogical weights will inevitably be present, whether recognized or not.

The data used in the preceding examples may be utilized to illustrate methods of weighting and to show the effects of varying weights upon the values of index numbers. The weights employed in constructing index numbers of farm crop prices may be either the quantities or values of the crops produced, depending upon the type of index selected. The quantities produced during the period 1910–1923 are given in the following table: [1]

[1] For the period 1910–19 the figures are taken from "An Index of the Physical Volume of Production," E. E. Day, 1921, 8. (Reprinted from the *Review of Economic Statistics*, Sept., 1920.) For the years 1920–23 the data are from *Weather, Crops and Markets.*

TABLE 50

Annual Physical Production, Twelve Crops, 1910–1923

Year	*Corn (Millions of bu.)*	*Cotton (Millions of bales)* [1]	*Hay, Tame (Millions of short tons)*	*Wheat (Millions of bu.)*	*Oats (Millions of bu.)*	*White Potatoes (Millions of bu.)*	*Sugar (Millions of lbs.)* [2]	*Barley (Millions of bu.)*	*Tobacco (Millions of lbs.)*	*Flaxseed (Millions of bu.)*	*Rye (Millions of bu.)*	*Rice (Millions of bu.)*
1910	2886	11.61	69.38	635.1	1186	349.0	3618	173.8	1103	12.72	34.90	24.51
1911	2531	15.69	54.92	621.3	922	292.7	4296	160.2	905	19.37	33.12	22.93
1912	3125	13.70	72.69	730.3	1418	420.6	3764	223.8	963	28.07	35.66	25.05
1913	2447	14.16	64.12	763.4	1122	331.5	3941	178.2	854	17.85	41.38	25.74
1914	2673	16.14	70.07	891.0	1141	409.9	4134	195.0	1035	13.75	42.78	23.65
1915	2995	11.19	85.92	1025.8	1549	359.7	4230	228.8	1062	14.03	54.05	28.95
1916	2567	11.45	91.19	636.3	1252	287.0	4671	182.3	1153	14.30	48.86	40.86
1917	3065	11.30	83.31	636.7	1593	442.1	3950	211.8	1249	9.16	62.93	37.74
1918	2503	12.04	76.66	921.4	1538	411.9	4220	256.2	1439	13.37	91.04	38.61
1919	2917	11.42	91.33	941.0	1248	357.9	3634	165.7	1390	8.92	88.48	41.06
1920	3209	13.44	87.85	833.0	1497	403.3	4665	189.3	1582	10.77	60.5	52.07
1921	3069	7.95	82.38	814.9	1078	361.7	....	154.9	1070	8.03	61.7	37.61
1922	2906	9.76	95.88	867.6	1216	453.4	....	182.1	1247	10.37	103.4	41.40
1923	3054	10.08	89.90	785.7	1230	412.4	....	198.2	1475	17.43	63.0	33.26

[1] Bales of 500 pounds gross weight.

[2] The figures for sugar production "represent total production of beet and cane sugar in contiguous United States for the crop year indicated, plus imports from non-contiguous United States for the fiscal year beginning July 1 of the crop year." For the years 1921–23 computations were based upon estimated values for this series.

Weighted Aggregates of Actual Prices

The thoroughly illogical results obtained when actual prices, as quoted, are totaled to secure an index number have been pointed out. The same objection cannot be made when the prices are appropriately weighted before the aggregate is taken. If for weights we employ the quantities produced in the base year (at time "0") the formula for the weighted aggregate is

$$\frac{\Sigma\, p_1\, q_0}{\Sigma\, p_0\, q_0}.$$

This is, in effect, the method employed by the United States Bureau of Labor Statistics, though the quantities are taken from a year other than the base year. The method is illustrated in the table to be found on page 208.

The desired index numbers, in the form of relatives, may be computed from the aggregates secured by totaling columns (5) and (8). Either year may be taken as the base, and the price aggregate in the other year expressed as a relative on this base. With the 1910 aggregate as base the index for 1911 is 111.5. Index numbers similarly computed for the other years are given in column (2), Table 54.

Another type of weighted aggregate may be constructed, with weights taken not from the base period but from the later period in the given comparison. That is, we may employ q_1 (quantity at time "1") as weight in comparing prices at time "1" with prices at time "0," and employ q_2 (quantity at time "2") as weight in comparing prices at time "2" with prices at time "0". Algebraically, the formula for the index number at time "1" is

$$\frac{\Sigma\, p_1\, q_1}{\Sigma\, p_0\, q_1}.$$

The process of computation is precisely the same as in the preceding example, except that the weights are changed

TABLE 51

Computation of Weighted Aggregates of Actual Prices

(1) *Commodity*	(2) *Unit*	(3) *Price, 1910*	(4) *Weight (Quantity produced 1910, in millions)*	(5) *Price × Weight* (p_0 q_0)	(6) *Price, 1911*	(7) *Weight (Quantity produced 1910, in millions)*	(8) *Price × Weight* (p_1 q_0)
Corn	bu.	$ .480	2,886	$1,385,280,000	$ .618	2,886	$1,783,548,000
Cotton	lb.	.141	5,805	818,505,000	.088	5,805	510,840,000
Hay	ton	12.14	69.38	842,273,200	14.29	69.38	991,440,200
Wheat	bu.	.883	635.1	560,793,300	.874	635.1	555,077,400
Oats	bu.	.344	1,186	407,984,000	.450	1,186	533,700,000
Potatoes (white)	bu.	.557	349.0	194,393,000	.799	349.0	278,851,000
Sugar	lb.	.0393	3,618	142,187,400	.0494	3,618	178,729,200
Barley	bu.	.578	173.8	100,456,400	.869	173.8	151,032,200
Tobacco	lb.	.093	1,103	102,579,000	.094	1,103	103,682,000
Flaxseed	bu.	2.317	12.72	29,472,240	1.821	12.72	23,163,120
Rye	bu.	.715	34.90	24,953,500	.832	34.90	29,036,800
Rice	bu.	.678	24.51	16,617,780	.797	24.51	19,534,470
				$4,625,494,820			$5,158,634,390

with each successive year. The index numbers secured by this method are given in column (3), Table 54.

The weights in these two cases have been *quantities*, for prices, multiplied by quantities, give aggregates in terms

of prices. But in weighting individual price relatives quantities will not serve. The abstract relatives must be weighted by *values*, if the resulting products are to be comparable. For values are in terms of a common dollar unit, while quantities may be expressed in a variety of units. The values which are to be employed as weights may be derived in various ways.

Fisher [1] outlines the four following methods, of which the second and third are hybrid types:

I. Each weight = base year price × base year quantity ($p_0 q_0$).
II. Each weight = base year price × given year quantity ($p_0 q_1$).
III. Each weight = given year price × base year quantity ($p_1 q_0$).
IV. Each weight = given year price × given year quantity ($p_1 q_1$).

Just as certain averages possess inherent bias, so a distinctive weight bias arises from each type of value weighting. (This inherent bias is absent from the quantity weighting.) A downward bias arises from weighting systems I and II (in which base year prices are used), while an upward bias arises from weighting systems III and IV (using prices in the given year). This is in part capable of mathematical demonstration [2] and has in part been established by numerous trials.

[1] *The Making of Index Numbers*, 54.

[2] An index weighted by type III must exceed an index weighted by type I. Weighting the price relative of a given commodity by type III, we have

$$\frac{p_1}{p_0} \times p_1 q_0$$

while by type I we have

$$\frac{p_1}{p_0} \times p_0 q_0$$

If p_1 exceeds p_0 (if the price relative is above 100) the weight by type III ($p_1 q_0$) is greater than the weight by type I ($p_0 q_0$). That is, all relatives above 100 are more heavily weighted by type III than by type I. But if p_1 is less than p_0 the weight by type III ($p_1 q_0$) is less than the weight by type I ($p_0 q_0$). All relatives below 100 are less heavily weighted by type III than by type I. Thus the effect of all price increases is over-emphasized and the effect of all price declines is underemphasized by type III, giving a net result always greater than type I. The same is true of type IV as compared with type II. As between types I and IV

In the several examples next following we shall deal only with values of quantities produced in the base year, 1910. These values are given in column (3) of Table 52. For weighting purposes they are taken to the nearest million.

Weighted Arithmetic Averages of Relative Prices

In the computation of an index of this type, each relative is multiplied by the appropriate weight and the sum of the products is divided by the sum of the weights. The process is illustrated in the following table.

Table 52

Computation of Weighted Arithmetic Averages of Relative Prices

(1) *Commodity*	(2) *Relative Price, 1910*	(3) *Weight*	(4) *Relative Price × Wgt.*	(5) *Relative Price, 1911*	(6) *Weight*	(7) *Relative Price × Wgt.*
Corn	100	$1,385	$138,500	128.8	$1,385	$178,388.0
Cotton	100	819	81,900	62.4	819	51,105.6
Hay	100	842	84,200	117.7	842	99,103.4
Wheat	100	561	56,100	99.0	561	55,539.0
Oats	100	408	40,800	130.0	408	53,407.2
White Potatoes	100	194	19,400	143.5	194	27,839.0
Sugar	100	142	14,200	125.6	142	17,835.2
Barley	100	100	10,000	150.2	100	15,020.0
Tobacco	100	103	10,300	101.1	103	10,413.3
Flaxseed	100	29	2,900	78.7	29	2,282.3
Rye	100	25	2,500	116.2	25	2,905.0
Rice	100	17	1,700	117.5	17	1,997.5
		$4,625	$462,500		$4,625	$515,835.5

(The weights employed are the values of the quantities produced in 1910, in millions.)

$$\text{Weighted arithmetic mean (1910)} = \frac{\$462{,}500}{\$4{,}625} = 100.$$

$$\text{Weighted arithmetic mean (1911)} = \frac{\$515{,}835.5}{\$4{,}625} = 111.5.$$

there is no necessary relation, but in general an index weighted by type IV will exceed an index weighted by type I. Base year weighting involves a downward bias while given year weighting involves an upward bias. (For a more detailed discussion of bias in weighting see Fisher, *The Making of Index Numbers*, Chapter V and pages 384–387.)

This value for 1911, it will be noted, is identical with that secured from the computations illustrated in Table 51. That index is a weighted aggregate of actual prices, the weights being the quantities produced in the base year. An arithmetic mean of relative prices, weighted by values in the base year, is always equal to a relative constructed from such an aggregate.[1]

Weighted Geometric Averages of Relative Prices

The process of computing the weighted geometric mean is identical with that of computing the unweighted geometric mean, except that the logarithm of each relative is multiplied by the given weight and the sum of these weighted logarithms is divided by the sum of the weights, the result being the logarithm of the desired index.[2] The method is illustrated in the following table:

[1] This may be readily demonstrated algebraically. The value of any commodity in the base year is p_0q_0, while the price relative for a second year is $\frac{p_1}{p_0}$. The weighted mean of such price relatives is equal to

$$\frac{\frac{p_1'}{p_0'} \times q_0'p_0' + \frac{p_1''}{p_0''} \times p_0''q_0'' + \frac{p_1'''}{p_0'''} \times p_0'''q_0''' + \ldots}{p_0'q_0' + p_0''q_0'' + p_0'''q_0''' + \ldots}$$

which reduces to

$$\frac{\Sigma p_1q_0}{\Sigma p_0q_0},$$

a weighted aggregate of the type mentioned.

In the same way the harmonic mean, weighted by full values in the second year, reduces to

$$\frac{\Sigma p_1q_1}{\Sigma p_0q_1}.$$

This has already been encountered as an aggregate of actual prices weighted by quantities in the second year.

[2] The formula for the weighted geometric mean is given on page 136 above.

TABLE 53

Computation of Weighted Geometric Average of Relative Prices, 1911

(1910 = 100)

Commodity	*Relative Price*, 1911	*Logarithm of Relative Price*	*Weight (value produced in* 1910)	*Logarithm of Relative Price* × *weight*
Corn	128.8	2.10992	1,385	2922.23920
Cotton	62.4	1.79518	819	1470.25242
Hay	117.7	2.07078	842	1743.59676
Wheat	99.0	1.99564	561	1119.55404
Oats	130.9	2.11694	408	863.71152
White Potatoes	143.5	2.15685	194	418.42890
Sugar	125.6	2.09899	142	298.05658
Barley	150.2	2.17667	100	217.66700
Tobacco	101.1	2.00475	103	206.48925
Flaxseed	78.7	1.89597	29	54.98313
Rye	116.2	2.06521	25	51.63025
Rice	117.5	2.07004	17	35.19068
			4,625	9401.79973

$$\text{Log } M_g = \frac{9401.79973}{4625} = 2.032822$$

$$M_g = 107.9$$

The index for 1911 on the 1910 base is 107.9. Values secured for all the years of the period covered are given in column (5), Table 54, together with the other weighted index numbers already explained.

How are we to judge of the relative merits of these three index numbers? We may, first, apply the time reversal test which was employed in comparing the five simple index numbers. This test is not met by any of the weighted types we have constructed. The geometric is equally at fault with the others. Though the simple geometric meets the test, the introduction of weighting imparts a bias to the result. Judged by that test alone none of the three is satisfactory. We may next try the second fundamental test which Fisher has developed, which is termed the "factor reversal test."

The Factor Reversal Test

The total value of any commodity in any year is, of course, the product of the quantity produced and the price per unit; algebraically, it is equal to $p_0' q_0'$. The ratio of the total value in one year to the total value in the preceding year is $\frac{p_1' q_1'}{p_0' q_0'}$. If, from one year to the next, both price and quantity should double, the price relative would be 200, the quantity relative 200 and the value relative 400. The total value in the second year would be four times the value in the first year. The value relative would be equal to the product of the price and quantity relatives, a relationship which is obvious in the case of a single commodity.

If, for a number of commodities, we construct an index of the price change from one year to the next and an index of the quantity change from one year to the next, we should expect their product to be equal to the ratio of the total values in the second year to the total values in the first year. If the product is not equal to the value ratio, there is an error in one or both of the index numbers.

As an illustration, we may apply this test to the first aggregative index constructed $\left(\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}\right)$. An index of quantities may be computed from this same formula, merely interchanging the q's and the p's; the formula becomes

$$\frac{\Sigma q_1 p_0}{\Sigma q_0 p_0}.$$

The same price factor appears in numerator and denominator, as we desire to measure only the effect of the quantity change. Substituting the given values of the twelve farm crops we have

$$\text{Quantity index, 1911 (1910 = 100)} = \frac{\$4,446,264,630}{\$4,625,494,820} = .96125.$$

In percentage form the index of quantities produced in

1911 is 96.125, with 1910 as base. The corresponding price index, by the same formula, is 111.5. The product

$$.96125 \times 1.115 = 1.0718.$$

That is, if prices have increased 11.5 per cent, while quantities have decreased 3.875 per cent, the total value should show an increase of 7.18 per cent.

For the value ratio we have

$$\frac{\Sigma\, p_1\, q_1}{\Sigma\, p_0\, q_0} = \frac{\$4,748,718,320}{\$4,625,494,820} = 1.02664.$$

There is a discrepancy here of $4\frac{1}{2}$ per cent. This formula does not meet the factor reversal test, and cannot be accepted as satisfactory.

When this test is applied to the second aggregative index we secure the following values for 1911, with respect to 1910 as base:

$$\text{Price index} = \frac{\Sigma\, p_1\, q_1}{\Sigma\, p_0\, q_1} = 106.8.$$

$$\text{Quantity index} = \frac{\Sigma\, q_1\, p_1}{\Sigma\, q_0\, p_1} = 92.05.$$

$$\text{Product} = .9205 \times 1.068 = .9831.$$

(In securing the product the index numbers are put in the ratio, not in the percentage form.)

Here is an error in the other direction of over four per cent.

The weighted geometric average also fails to meet this fundamental factor reversal test. With respect to both the geometric index and the aggregates we have, apparently, by the introduction of weights spoiled index numbers which in their simple form were unbiased. Yet weights we must have, if the index numbers are to represent the facts accurately. Neither a simple index nor a weighted form of a simple index will meet the two tests laid down as

fundamental. Professor Fisher tested 46 such formulas, of which only four (the simple geometric, median, mode and aggregative) met the time reversal test, and none met the factor reversal test.

The "Ideal" Index

A way out of this difficulty is offered by the possibility of "rectifying" formulas by a crossing process, by averaging geometrically formulas which err in opposite directions. Professor Fisher has made exhaustive trials of all possible formulas by this process, finding thirteen formulas in all which met both tests. Of these he has selected one as "ideal," from the viewpoint of both accuracy and simplicity of calculation. This ideal index is the geometric mean of the two aggregative types illustrated above. Its formula [1] is

$$\sqrt{\frac{\Sigma\, p_1\, q_0}{\Sigma\, p_0\, q_0} \times \frac{\Sigma\, p_1\, q_1}{\Sigma\, p_0\, q_1}}.$$

This index may be computed readily, in the present instance, from the results already obtained. Thus for 1911 we have

$$\begin{aligned} \text{Ideal index} &= \sqrt{111.5 \times 106.8} \\ &= 109.14 \end{aligned}$$

This index number meets both the time reversal and the factor reversal test. Applying the former:

$$\begin{aligned} &\text{Index of prices, 1911 (1910 = 100)} = 109.14 \\ &\text{Index of prices, 1910 (1911 = 100)} = 91.63 \\ &1.0914 \times .9163 = 1.00 \end{aligned}$$

For the factor reversal test, applied to the data for 1911, (with 1910 as base) we have

$$\text{Index of prices} = \sqrt{\frac{\Sigma\, p_1\, q_0}{\Sigma\, p_0\, q_0} \times \frac{\Sigma\, p_1\, q_1}{\Sigma\, p_0\, q_1}} = 109.14.$$

[1] The same formula was developed independently by Bowley, Pigou, Walsh and Young. See *The Making of Index Numbers*, xv, 240–242.

$$\text{Index of quantities} = \sqrt{\frac{\Sigma\, q_1\, p_0}{\Sigma\, q_0\, p_0} \times \frac{\Sigma\, q_1\, p_1}{\Sigma\, q_0\, p_1}} = 94.07.$$

$$\text{Index of values} = \frac{\Sigma\, p_1\, q_1}{\Sigma\, p_0\, q_0} = 102.664.$$

Product of price and quantity indices = $1.0914 \times .9407 = 1.02668$

The ideal index, the two weighted aggregates which enter into its construction and the geometric mean weighted by values in the base year are given in the following table, for the years 1910 to 1923. The index numbers are plotted in Fig. 52.

TABLE 54

Comparison of Weighted Index Numbers of Farm Crop Prices, 1910–1923 [1]

(1) *Year*	(2) *Aggregative (Weighted by base year quantities)* $\frac{\Sigma\, p_1 q_0}{\Sigma\, p_0 q_0}$	(3) *Aggregative (Weighted by given year quantities)* $\frac{\Sigma\, p_1 q_1}{\Sigma\, p_0 q_1}$	(4) *Ideal index Geometric mean of indices in cols. (2) & (3)*	(5) *Weighted Geometric Average (Weighted by base year values)*
1910	100.0	100.0	100.0	100.0
1911	111.5	106.8	109.1	107.9
1912	94.4	93.7	94.0	94.0
1913	112.4	109.4	110.9	109.9
1914	102.8	98.9	100.8	97.0
1915	102.3	102.3	102.3	101.0
1916	156.8	151.3	154.0	151.2
1917	212.5	210.9	211.7	207.2
1918	224.2	220.3	222.2	219.3
1919	241.3	238.2	239.7	235.1
1920	139.9	137.9	138.9	137.2
1921	104.3	103.6	104.0	101.6
1922	131.6	127.9	129.7	128.9
1923	149.3	145.2	147.2	144.1

The wide discrepancies which were found between the various simple index numbers do not appear when the

[1] Values for the two sets of aggregative index numbers for the years 1910 to 1919, inclusive, were computed by W. M. Persons. ("Fisher's Formula for Index Numbers," *Review of Economic Statistics*, Prel. Vol. III, 107.)

weighted indices are compared. There are significant differences, but there is none of the erratic behavior of some of the simpler forms.

Of these four types the "ideal" index undoubtedly serves as the best measure of the average price change between 1910 and each of the given years. It is designed, it should

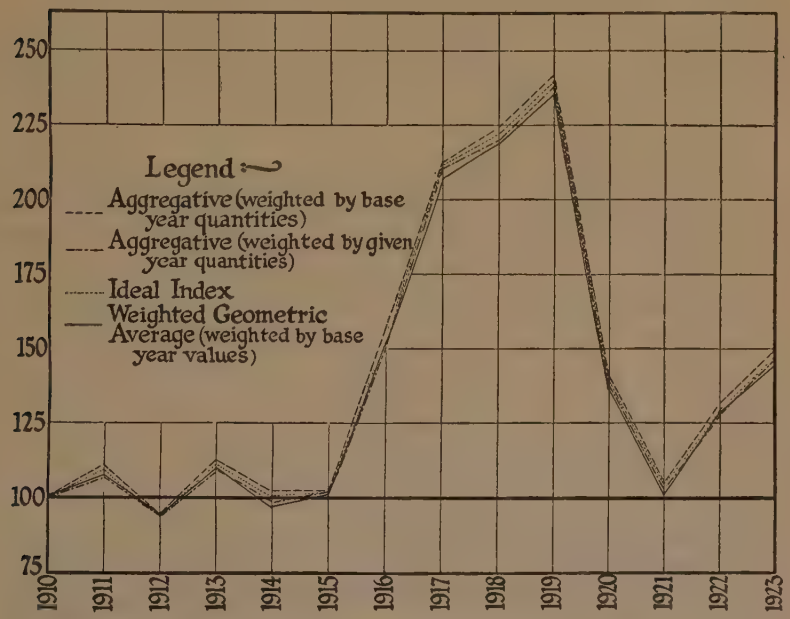


FIG. 52. — Comparison of Four Weighted Index Numbers of Farm Crop Prices 1910–1923. (1910 = 100)

be remembered, to measure the change between two stated times, and not for intermediate comparison. The value of the index for 1923, for instance, is determined by the relation between prices and quantities in 1910 and in 1923. There is double weighting and the weights vary from year to year. If 1923 is to be compared with 1922 a new index is needed, in which the prices and quantities for 1923 and 1922 alone are included. Direct comparison on the basis of the values for the "ideal" index given in the above

table is liable to error, because of the weighting system employed.

It is one of the merits of the geometric mean with constant weights that it permits the index for each year to be compared directly not only with the base year index, but with the index for any other year. The base may be shifted directly from the relatives, and the same result will be secured as if the computation were made from the original data. If this same system be followed with the "ideal" index no large errors may be expected, but strict accuracy will not be secured.[1]

The chief obstacles in the way of general adoption of the ideal index arise from the difficulty of obtaining annual or monthly quantities to use as weights, and from the time involved in its computation. Where accuracy is essential the latter is not a serious difficulty. As a substitute formula which is much more quickly calculated Fisher has proposed

$$\frac{\Sigma\ (q_0 + q_1)\ p_1}{\Sigma\ (q_0 + q_1)\ p_0}.$$

This formula, which has also been recommended by Edgeworth and Marshall, is considered by Fisher to be "the best practical all-around formula, taking all four points into account — accuracy, speed, minimum legitimate circular discrepancy, simplicity." Results from this formula will generally differ from those secured from the "ideal" formula by less than one fourth of one per cent. The following table illustrates the method of computation, data for 1910 and 1911 being employed.

[1] If year to year comparison be a primary aim in a given instance, the "ideal" index may be constructed on the chain system. Link index numbers are first constructed, each year serving as base for the computation of the index for the succeeding year. These links may then be "chained" with reference to a fixed base. Warren M. Persons has shown that the errors involved in following this method are cumulative, and may be serious if the links are chained for a number of years.

TABLE 55

Computation of Aggregative Index, Weighted by Combined Quantities

(1) *Commodity*	(2) *Unit*	(3) *Price, 1910*	(4) *Quantity 1910 + quantity 1911 (in millions)*	(5) *Price 1910 × sum of quantities col. (3) × col. (4)*	(6) *Price, 1911*	(7) *Price 1911 × sum of quantities col. (6) × col. (4)*
Corn.....	bu.	$ .480	5,417	$2,600,160,000	$ .618	$3,347,706,100
Cotton ...	lb.	.141	13,650	1,924,650,000	.088	1,201,200,000
Hay.....	ton	12.14	124.30	1,509,002,000	14.29	1,776,247,000
Wheat....	bu.	.883	1,256.4	1,109,401,200	.874	1,098,093,600
Oats.....	bu.	.344	2,108	725,152,000	.450	948,600,000
Potatoes..	bu.	.557	641.7	357,426,900	.799	512,718,300
Sugar	lb.	.0393	7,914	311,020,200	.0494	390,951,600
Barley....	bu.	.578	334	193,052,000	.869	290,246,000
Tobacco..	lb.	.093	2,008	186,744,000	.094	188,752,000
Flaxseed..	bu.	2.317	32.09	74,352,530	1.821	58,435,890
Rye.....	bu.	.715	68.02	48,634,300	.832	56,592,640
Rice.....	bu.	.678	47.44	32,164,320	.797	37,809,680
				$9,071,759,450		$9,907,352,710

$$\frac{\Sigma(q_0 + q_1)p_1}{\Sigma(q_0 + q_1)p_0} = \frac{\$9,907,352,710}{\$9,071,759,450}$$

$$= 109.2 \text{ (index for 1911 on 1910 base).}$$

(The index is here expressed in percentage form.)

This formula requires the same data as the "ideal" index, and these are not generally to be had. Usually it is only possible to secure comprehensive quantity figures at each census period, and for the intervening years constant weights must be employed. In such cases the weighted aggregative

$$\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}$$

is probably the most generally useful type. The weighted geometric has many virtues, but is subject to a definite weighting bias. If no weights can be secured, or even approximated, the simple geometric and the simple median are far better than any of the other simple types. The geometric mean is more generally useful than the median.

The Reliability of Different Index Numbers

An index number of prices is always based upon the study of a sample, the result being taken as representative of the entire field of prices from which the particular sample was drawn. Some method is needed, therefore, by which we may judge of the reliability of the different types of index numbers, of their probable stability when computed from a number of successive samples. Some differences might be expected between index numbers based upon different samples. With which type of index number would these differences due to fluctuations of sampling be least? [1]

Truman L. Kelley [2] has attempted to measure the probable errors of the chief types of index numbers and has graded these types on the basis of excellence in this respect. Two index numbers, the weighted geometric mean and the weighted median, are given the highest grade, as being the most reliable, the least affected by fluctuations of sampling. Fisher's "ideal" index is ranked somewhat lower, though above the weighted arithmetic and harmonic averages of price relatives. The simple unweighted arithmetic average of relatives is given the lowest rating in the list.

For reliability, flexibility and general excellence Kelley selects the weighted geometric mean as the best type of price index number. A ratio of aggregates

$$\frac{\Sigma(p_1\, w)}{\Sigma(p_0\, w)}$$

with selected weights (not necessarily precisely equal to the quantities marketed or consumed) is given a total score, based on the essential requirements of a good index number, as high as that of the weighted geometric mean and higher than that of the "ideal" index. Weights other than actual

[1] The subject of sampling, in relation to the reliability of statistical measures, is discussed in greater detail below.

[2] *Statistical Method*, 334–346.

quantities are used in order that there may be flexibility in the matter of weighting.

Other Problems Involved in the Construction of Price Index Numbers

The preceding section has dealt with the technical problems connected with the averaging of a given set of data in order to secure an index number of price variations. Certain methods have been shown to be quite faulty, while certain others have been found to be appropriate for given purposes. One who would use index numbers with intelligence should understand fully the methods which have been employed in securing given results, in order that he may know precisely what the given figure is designed to measure and what degree of reliability attaches to it.

Such problems as these are not the only ones which confront those who construct index numbers, nor are these considerations the only ones which users of index numbers should bear in mind. Of equal importance with problems of averaging and weighting are the practical questions connected with the selection of representative samples. The only completely accurate measure of the general level of commodity prices would be secured by determining the ratio between all money units (including credit) in circulation and all the physical units of goods exchanged for money over a given period. The measurement of general price changes between two periods would thus involve complete knowledge of these two factors for each of the two periods. Such knowledge, of course, can not be had, so recourse must be had to the method of sampling. And primary importance attaches to the number of commodities and the character of the commodities upon the prices of which a given index number is based.

Number of Commodities to be Included

Here again we are confronted with a relation which has already been mentioned, the relation between methods and uses. Decision as to the number of commodities and the kinds of commodities to be included in a given case must rest upon the purpose for which the index is to be constructed. Assuming that the index number is to serve as a measure of general changes in the price level, the question as to the number of commodities to be included may be easily answered — the larger the sample the more representative will be the results. The frequency polygon based upon a large sample will approach more closely to the ideal curve which would represent all price quotations than will that based upon a small sample. Thus, as a measure of general price changes, more confidence may be placed in the Bureau of Labor Statistics index, which is based upon 404 price quotations, than in Bradstreet's, which is based upon 96 quotations, though the latter has particular virtues of its own. Yet index numbers based upon a small number of quotations may not be ruled out as without value. Wesley C. Mitchell, whose researches have materially increased our knowledge of the price system and of the characteristics of index numbers, has compared in detail index numbers based upon varying numbers of quotations. Unexpected similarities are found. Those constructed from a limited number of quotations reflect the broad movements of prices in much the same way as do those based upon the prices of several hundred commodities. In important details there are differences, however, differences which may involve doubt as to the movement of prices in a given year. In such cases the index numbers based upon many quotations must be accepted as the more representative of general price movements, provided that the commodities included be equally representative of the various elements in the price system.

For other purposes, however, index numbers based upon a limited number of quotations may be preferable. This is particularly true when a "sensitive" index is desired, one which will serve as a forecaster of general price movements rather than as a precise measure of changes in the general price level. Of this type is the index constructed by the Federal Reserve Bank of New York, formerly based upon quotations of 12 basic commodities (raw materials), and now upon quotations of 20 commodities. Somewhat similar in character is the index of wholesale prices constructed by Warren M. Persons, which is based upon only 10 price series. The object in this case has been to secure a price index of business cycles, a measure of the swings of business through phases of prosperity and depression. When a sensitive index is required the object is attained by the selection of a limited number of commodities the prices of which are subject to extreme fluctuations, rather than by the inclusion of a great many commodities. Yet the uses to which an index of this type may be put are limited. The "sluggishness" of the many-commodities index number is a sluggishness which inheres in the price system, and which must be reflected in a faithful index of general prices.

The question of the number of commodities to be included can not be discussed apart from that of the character of the commodities upon the quotations of which an index number is based. The representative character of an index number rests in part upon the number of price series included, but the nature of these series is of even greater importance. In the selection of price quotations the fact must be recognized that there are characteristic differences in the behavior of the prices of different commodity groups. These groups of prices, their inter-relations, their behavior, their relation to the functioning of the economic system and to the swings of prosperity and depression, are matters of immediate and practical importance to economists and business men.

Elements of the Price System

It is impossible to give, by graphic means, a true conception of the complicated inter-relations which prevail within the price system. In Fig. 53 certain of the broader relations only are brought out and some of the more important

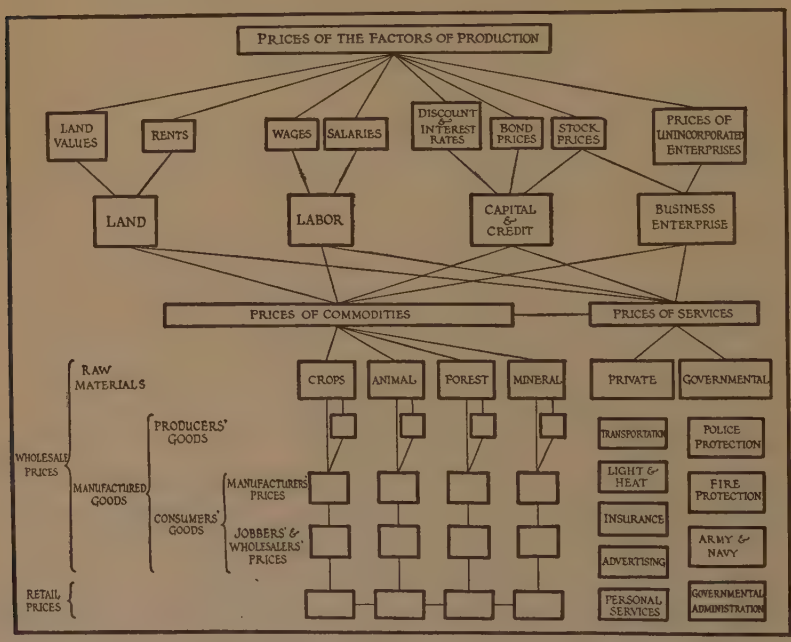


Fig. 53. — A Graphic Representation of the Relation between Certain Elements of the Price System

elements are indicated. The broad subdivisions are prices of the factors of production, prices of commodities at wholesale, prices of commodities at retail, and prices of services, including those rendered by private, profit-seeking enterprises and those rendered by governmental units. The prices of private services come in part within the category of wholesale prices, and in part, in so far as they represent final services to ultimate consumers, within that of retail

prices. Within each of these subdivisions there are important groups of related prices. Within the first subdivision, that of the prices of the factors of production, it is found that rents and land values, salaries and wages, discount rates and interest rates, bond prices and stock prices, all have their characteristic modes of behavior during the upward and downward swings of prices. In the field of commodity prices the same differences are found and, again, the various services rendered by private and public bodies are characterized by price changes peculiar to themselves.

All the lines of relationship existing between the different units and groups in the price system can not be indicated graphically. Were this possible, lines would be seen to radiate from every unit and every group to every other unit and group. Certain of these relationships are direct and significant, while others are tenuous, impossible to trace in practice.

Nor is the representation of the price system by a plane surface a just one. This picture shows it at a moment of time. Stretching backward are countless ties, connecting every price with preceding prices, and reaching forward are more ties, for future prices and price relationships derive directly from present prices. Thus, for example, it is obvious that the prices of manufactured goods today have been affected by the price of raw materials, labor, etc., in the past.

An accurate index of the purchasing power of money or an accurate measure of changes in the general price level should include representatives of all these groups and subgroups within the price system. No index currently published does this. Certain index numbers of wholesale prices purport to serve as measures of general price changes, but they can not be accepted as such.

We are at present concerned with the construction of index numbers of wholesale prices. What has been said should serve to emphasize the true scope and significance of

such index numbers and to indicate that they measure changes in but one element, though perhaps the most important, in the price system. It remains to consider some of the different groups of related prices within this division of the price system with particular reference to the problem of index number construction.

Price Groups in the Field of Wholesale Prices

Since an index number of wholesale prices must rest upon sample quotations, the sample must be representative, must include commodities whose prices are typical of the various elements in the price system. The division into elements for this purpose must be based upon the character of the price changes peculiar to the different groups. Of the groups thus distinguished, the most obvious are those representing different industries. Textile prices and steel prices, leather prices and the prices of chemicals are subject to different influences. Trade depressions and revivals do not affect all industries at the same time or in the same way, so that an index of wholesale prices must include quotations from all important industrial groups. If preponderant influence upon an index is exerted by the prices of certain types of commodities, the index, by that much, loses its representative character. Thus Bradstreet's index, it has been established, gives greater weight to cotton fabrics, hides and leather, and cured meats than is justified by their actual importance in trade, a fact which does not detract from its utility for some purposes but which lessens its value as a representative index of wholesale prices.

The extent of these differences between the price movements of commodities in different industrial groups may be appreciated by comparison of the index numbers of wholesale prices of farm products and house furnishing goods since 1913.

In order that an index may be representative it is not alone sufficient that all industries be given an appropriate number of representatives in the sample. Raw materials and manufactured goods show characteristic differences in their fluctuations, and fitting representation must be given to each of these groups. Prices of the former are, in general, more sensitive to changes in business conditions, their movements preceding those of manufactured goods and showing more violent fluctuations. There are two related reasons for these facts. Raw materials are traded in for purposes of manufacture and sale. When business improves after a period of depression, increased demand on the part of consumers (or expected increase in demand) leads competing manufacturers to bid against each other for raw materials. It is in the raw material markets that the pressure of increased demand first centers, and this bidding causes prices to rise in these markets before the prices of other goods are affected. (This is not an invariable rule, as was demonstrated during the revival of 1919.) Similarly, a period of crisis and liquidation begins with large stocks of manufactured goods on hand, and at the first evidence of slackening trade demand for raw materials falls off. Business forces pure and simple play in the raw material markets with more freedom than in the markets for manufactured goods. Hence the tendency of prices in these markets to anticipate, in their movements, prices in other commodity markets.

The second reason for the greater stability of prices of manufactured goods is found in the fact that these prices include a greater percentage of stable cost factors, namely, overhead charges and labor costs. Wages, interest, rents move more slowly and less violently than do commodity prices. The inclusion of these elements in commodity prices tends to render these prices more stable. Therefore, as commodities move forward from the raw stage to their final manufactured condition their prices include more and

more of these stabilizing elements, and become less violent in their fluctuations.[1] Important differences between different classes of manufactured goods arise from this fact.

Each of the groups last mentioned contains minor groups of commodities with distinct price characteristics. Within the raw material group there are marked differences between agricultural products, animal products, forest products and mineral products with respect to price movements. Agricultural products are affected by weather and crop conditions as well as by business conditions, and, though subject to price fluctuations of some magnitude, reflect prevailing business conditions less accurately than do the prices of mineral products.[2] Animal and forest products appear to stand between these two with respect to the faithfulness with which they reflect business conditions in their price movements. Thus, in selecting raw materials for inclusion in a sample of price quotations from which a representative index number is to be constructed, fair weight must be given to these various classes.[3]

Manufactured goods, again, do not constitute a single homogeneous group with respect to their price movements. In so far as they are to be used for further production, or to undergo further manufacture, they resemble raw materials in relation to the bidding of competing manufacturers, and their prices, therefore, are characterized by relatively wide oscillations. In so far as the demand for them is for the purpose of final consumption, purely business forces have less weight and their prices are more stable. Related to this

[1] Cf. Mitchell, "The Making and Using of Index Numbers" (*Bulletin 284*, U.S. Bureau of Labor Statistics), 44–45, for examples.

[2] It should not be inferred from this that there is no relation between agricultural production and the prices of agricultural products, and general business conditions. Good reasons have been advanced for believing that there is a direct causal relation between agricultural and business conditions. The immediate price relation, however, is frequently one of contradictory movements.

[3] Cf. "The Making and Using of Index Numbers," 47.

argument is that which has already been presented, the increasing stability of prices as the stable elements of wages and overhead charges bulk larger in commodity costs. So, again, the sample price quotations from which an index of wholesale prices is to be constructed must include prices representative of producers' and consumers' goods, of goods in the intermediate as well as the final stages of manufacture.[1]

Other divisions of the price system exist, but those indicated above are the most important from our present point of view. A representative index number of wholesale prices should be based upon price quotations drawn from all the groups indicated, with weight given to each group in proportion to the relative importance in trade of the commodities in that group.

AMERICAN INDEX NUMBERS OF WHOLESALE PRICES[2]

Index Numbers of the United States Bureau of Labor Statistics

The authoritative index of wholesale prices in the United States is that compiled by the United States Bureau of Labor Statistics. This index was first constructed in 1902, and was continued until 1913 as an unweighted average of relative prices, the base of each relative being the average price of the given commodity during the ten year period 1890–1899. The report for the year 1914 marked a material change in method, with a corresponding change in the base period employed in preparing the index for publication. As it at present stands the index for any given period

[1] *Ibid.*, 45, 49.

[2] Detailed descriptions and comparisons of these index numbers will be found in *Bulletin 284*, U.S. Bureau of Labor Statistics, *Index Numbers of Wholesale Prices in the United States and Foreign Countries.* The current bulletin of the Bureau of Labor Statistics on "Wholesale Prices" gives detailed information as to the character of the wholesale price index of that department.

(month or year) is a weighted aggregate of actual prices, the aggregate being expressed, to facilitate comparison, as a relative with 1913 as the base.

The index is now based upon 404 price series. (A single commodity may be represented by several quotations, the prices for different grades or in different markets being given. Thus for raw cotton there are two quotations, Middling, New Orleans, and Middling Upland, New York.) The price of each commodity is multiplied by a constant weight, the quantity marketed in 1919.

The following illustrates the method as applied to cotton:

Commodity	*Average Price,* 1919 (per lb.)	*Quantity marketed in* 1919	*Average Price,* 1919 × *Quantity marketed in* 1919
Cotton, Middling, New Orleans.....	$.319	3,806,921,000 lbs.	$1,214,407,799
Cotton, Middling Upland, New York..	.325	1,903,461,000 lbs.	618,624,825

When this process is carried out for the entire 404 price series included, the sum of the values in the last column gives the index number for the given period, in this case the year 1919. As published, this sum is expressed as a relative, the sum in 1913 representing 100. The method of construction renders it possible to shift the base to any desired year or month, changing the given relatives to percentages on the new base.

This index number, therefore, is based upon the cost at wholesale of a bill of goods. The bill of goods remaining the same, the total cost changes as the prices of the various commodities change, and the index measures the effect of these changing individual prices upon the total cost. Subtotals are found for nine commodity groups, and index numbers for these groups are published, as well as for the

complete list of commodities. The commodity groups are as follows: [1]

Farm products.
Foods.
Cloths and clothing.
Fuel and lighting.
Metals and metal products.
Building materials.
Chemicals and drugs.
House furnishing goods.
Miscellaneous.

Index Numbers of the Federal Reserve Board

The Federal Reserve Board publishes two different sets of index numbers of American wholesale prices. The first of these is based upon the data compiled by the Bureau of Labor Statistics, regrouped in order that certain elements of the price system may be more effectively studied. The weights and methods employed are those of the Bureau of Labor Statistics. The following are the groups for which index numbers are compiled by this Board:

Raw materials.
  Crops.
  Animal products.
  Forest products.
  Mineral products.
Producers' goods.
Consumers' goods.

The utility of this classification was first demonstrated by W. C. Mitchell. Its significance has been explained in the preceding pages. The group index numbers were first published in the *Federal Reserve Bulletin* for October, 1918,

[1] For lists of the commodities included in the several groups, see the latest bulletin on "Wholesale Prices," U.S. Bureau of Labor Statistics.

which also contained a list of the articles in each commodity group.

For the purpose of facilitating international comparison of price movements the Federal Reserve Board has begun the construction of a series of index numbers of wholesale prices in the chief commercial countries. This work was undertaken because of the wide variations in method and in the character of the commodities included in the construction of price index numbers in different countries, variations which made comparison difficult. At present these international price index numbers are being currently published for the United States, Canada, England, France and Japan.

The general method of securing these index numbers is that which has been described above in connection with the index of the Bureau of Labor Statistics. Money aggregates are secured by multiplying the price of each commodity by an appropriate weight, and getting the total of the individual products. These money aggregates are expressed as relatives, the aggregate for the year 1913 representing 100. The indices differ from that of the Bureau of Labor Statistics in the number of commodities included, in the grouping system employed and in the weights adopted.

Each of the international price indices is based upon 90 to 100 price quotations, representing about 70 different commodities. A double system of classification is employed. One system of grouping follows that already explained, a division into raw materials, producers' goods and consumers' goods. The same commodities are again classified into "goods produced" and "goods imported." Finally, an index of "goods exported" is constructed. The commodities, the prices of which are used in this index, are drawn from the groups previously mentioned, those of importance in the export trade being selected. These six group index numbers, together with the index of all commodities, are published for the various countries. An additional index in

which prices are corrected to the gold basis is also published for each of the foreign countries for which index numbers are constructed.

The weights employed in the construction of the group index numbers are the quantities produced, imported or exported in the year 1913. In securing the "all commodities" index the money aggregate for "goods produced" is added to that for "goods imported."

These index numbers have been constructed for the year 1913, and by months from January, 1919, to date (for France, since January, 1920).[1]

Index Numbers of the War Industries Board

The most comprehensive index of wholesale prices ever compiled in the United States was that constructed by the Price Section of the War Industries Board in connection with its studies of prices during the war. The period 1913–1918, inclusive, was covered by these studies. The results are shown in great detail in the *History of Prices During the War*, published by the War Industries Board, and are summarized in *Price Bulletin No. 1* of that series.

Price quotations were secured by the staff of this Board for 1474 commodities during the stated period. In the construction of index numbers 1366 of these price series were utilized. These commodities were grouped in fifty commodity classes and seven industrial groups. Index numbers were worked out for the fifty classes, the seven main groups and for "all commodities," as well as for many minor divisions of the classes.

These index numbers are all weighted aggregates of actual prices, expressed as relatives, a type similar to that of the United States Bureau of Labor Statistics. The weights consist of the quantities produced in the United States in

[1] For a detailed description of these index numbers see the pamphlet, *Prices in the United States and Abroad*, Federal Reserve Board, 1924.

the calendar year 1917, plus the quantities imported in that year. Actual prices for a given month, for the commodities within a given class, multiplied by the proper weights, and totaled, give the money aggregate for that month for the particular class being studied. In combining the class aggregates to form index numbers for the seven main groups and for all commodities, each class was weighted in accordance with its relative importance, precisely the same method being employed as in weighting the various commodities.

In converting these aggregates to relatives the base selected was the twelve month period, July, 1913, to June, 1914, the twelve months preceding the outbreak of war. The base may be shifted at will, however, without loss of accuracy.

In addition to the index numbers published by these governmental bodies there are several index numbers of wholesale prices compiled and published by private agencies. Of these, Bradstreet's and Dun's index numbers are the best known.

Bradstreet's Index of Wholesale Prices

Bradstreet's index is published as a sum of actual prices. It is constructed by reducing to a "per pound" basis the prices of 96 staple articles of commerce, and securing the total of these "per pound" prices. No system of weighting is employed. The monthly index is constructed from prices on the first day of each month, not average monthly prices, as in the case of the Bureau of Labor Statistics index. Since the prices are not expressed as relatives, there is no base period. The aggregate prices published may be expressed as relatives on any month or year as a base. The index covers the period 1892 to date.

It has been pointed out in an earlier section that all index numbers are weighted, whether the process be a conscious or an unconscious one. Bradstreet's index in practice gives

great weight to certain classes of commodities. In a list published in Bradstreet's on July 10, 1897, the per pound price of coke was $0.0007 while that of alcohol was $0.34 and of wool $.50. It is obvious that much greater weight is given each of the latter commodities when prices per pound are totaled. Warren M. Persons has worked out the following table showing the constitution of Bradstreet's index on September 1, 1921.[1]

TABLE 56

Distribution of Weights, Bradstreet's Index on September 1, 1921

Class of Commodity	*Weight (Percentage Distribution)*
Textiles, raw and manufactured:	
Cotton fabrics	16.7
Wool, raw	5.6
Other	3.3
Provisions and Groceries:	
Milk, eggs, butter and cheese	8.6
Cured pork and beef, and lard	6.9
Miscellaneous provisions	12.4
Hides and leather	13.3
Building materials	1.5
Coal and Coke:	
Coal	*
Coke	*
Metals:	
Iron and steel	.7
Other metals	4.3
Chemicals and drugs	9.6
Live stock	3.6
Oils	4.0
Fruits	3.2
Breadstuffs	1.0
Naval stores	.9
Miscellaneous	4.2

* Less than one tenth of one per cent.

This distribution of weights gives great influence to cotton fabrics, hides and leather, chemicals and drugs and cured meats, and the characteristics of this index number are due in large part to this weighting. The weights, moreover,

[1] *Review of Economic Statistics*, Prel. Vol. III, 365–366.

are not constant, for changing prices cause changes in the distribution of weights. Thus on February 1, 1920, textiles had a weight of 34.46% of the total. In the course of the ensuing eighteen months textile prices fell much more rapidly than general prices, so that by September 1, 1921, textiles had a weight of only 25.58% of the total.

In spite of this illogical weighting system Bradstreet's index has merit, and for certain purposes, particularly as a barometer of business conditions, is one of the most useful index numbers of wholesale prices.

Dun's Index of Wholesale Prices

The index number of wholesale prices published by the mercantile agency of R. G. Dun and Co., and appearing monthly in *Dun's Review*, resembles Bradstreet's index in that it is presented as a sum of actual prices, with no base period. The method employed in its construction, the distribution of weights and the periods covered by the two index numbers differ materially.

Dun's index is a statement, in dollars and cents, of the cost of a year's supply, for a single individual, of certain staple commodities. The monthly index is secured by multiplying the price of each of the staples included by the average annual per capita consumption of that commodity. Thus the index on January 1, 1921, was $198.600, which was the total cost, at prices prevailing on January 1, 1921, of a year's supply, for a single individual, of all the commodities included. The index is based upon prices prevailing on the first day of each month. The prices of approximately 300 commodities are utilized in its construction. No list of the actual commodities included has been published, nor has the exact number ever been stated. Quotations for only the necessities of life are utilized, luxuries being excluded. Though the weights are based upon per capita consumption, the index does not purport to be a measure of the cost of living. Staples such as pig iron, coal, building

materials, etc., which are not consumed directly by individuals, are included. This index, then, is designed to serve as a measure of general changes in the level of wholesale prices. It was first published in 1901, but the calculations have been carried back to 1860.

Logically, Dun's method of weighting each price by the average per capita consumption of the given commodity is excellent, but doubt arises as to how effectively the method is employed in practice. Adequate data concerning the consumption of commodities in the United States are lacking, except for a relatively few articles. Moreover, it is known from the published figures that the weight given to food is approximately 50 per cent of the total, a figure which is believed to be somewhat excessive in the light of such pertinent facts as are available. This excessive weighting of food accounts for certain of the characteristic movements of Dun's index, and explains its failure to accord closely at all times with the other general index numbers.

Fisher's Weekly Index Number of Wholesale Prices

In January, 1923, Irving Fisher launched a new index number of wholesale prices, computed and published weekly. This is the first comprehensive price index to appear on a weekly basis.

The index numbers are published as relative prices, on the 1913 base, but the relatives are derived from weighted aggregates of actual prices. The weights, which are constant, are the quantities marketed in 1919, as recorded by the Census Bureau. The formula employed is, therefore, the first aggregative described above

$$\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0},$$

except that weights are drawn from 1919 instead of the base year. The procedure is substantially that followed by the

United States Bureau of Labor Statistics. Fisher considers this to be the closest approach to the "ideal" index which the data permit.

In the matter of weighting this index number marks an interesting innovation. The index is based upon prices of 205 commodities, drawn from *Dun's Review*. For the most important commodities the actual quantities for the year 1919 are used as weights, with one slight correction to provide an appropriate class weight. To facilitate computation the weight given to each of the other commodities is one of the following numbers: 1000, 100, 10, 1, .1. In each case the round number nearest (geometrically) to the actual quantity is used as weight. This procedure was adopted after it had been established that the use of round weights for the less important commodities involved an error of less than one per cent.

The base aggregate of Fisher's index in the year 1923 was that of November, 1922, but the weekly index has been spliced on to the index of the Bureau of Labor Statistics. For November, 1922, the index of the Bureau of Labor Statistics was 156. In computing the 1923 indices the ratio of the aggregate for a given week to the base aggregate (November, 1922) was multiplied by 156. The present base aggregate is that of November, 1923, but a similar splicing process permits the index to appear as a relative, with 1913 as base.

The index for each week is published in the daily press on Monday of the following week. This promptness of publication gives an added value to the index, which is one of the most useful general price index numbers currently compiled.

Persons' Commodity Price Index of Business Cycles

The index numbers which have been described above have all been designed to serve as measures of changes in

the general level of prices, though their real significance is limited to the field of wholesale commodity prices. Several price index numbers have been constructed for special purposes, and it is probable that the future will see an increasing number of such restricted purpose indices. Of these, one of the most interesting is the "commodity price index of business cycles" constructed by Warren M. Persons and Eunice S. Coyle.[1]

The purpose for which this index has been constructed is "to measure changes in general business conditions during alternating periods of prosperity and depression." Otherwise stated, the problem is "to select and combine series of wholesale prices of commodities in order to secure an index of business cycles." The business cycle is primarily a price phenomenon. Other factors are of great importance, but all other forces are felt through their influence upon the price system, and it is through price changes that the cycle is brought home most immediately to business men, employees and consumers. While all elements in the price system are affected, security prices and wholesale commodity prices feel the influence of cyclical changes most directly and violently. Professor Persons has chosen to construct a wholesale price index which will reflect and measure such price movements as are directly related to the business cycle. His problem, then, is fundamentally different from that of makers of "general purpose" index numbers, and his methods are correspondingly different.

In determining the constitution of this index tests were applied to a number of price indices for industrial groups and to individual price quotations for a number of commodities, the object being the selection of those price series the movements of which corresponded with "typical cyclical fluctuations" of prices and general business during the period 1903–1914. Earlier investigations had shown that the cyclical movements of the price index numbers of the

[1] Cf. *Review of Economic Statistics*, Prel. Vol. III, 353–369.

Bureau of Labor Statistics and of Bradstreet's, as well as many non-price series, showed a general agreement during this period, crests of the wave movements coming in 1907, 1910 and 1912. The wave movement which was common to these various series was taken as the standard, and was employed in testing various price series in order that those with non-typical fluctuations might be eliminated.

Index numbers of prices for various industrial groups were first constructed, and these were tested in the manner described. It was established that price index numbers for mining, and for manufactures of iron and steel, leather, chemicals and textiles showed typical cyclical fluctuations; their movements corresponded with the cyclical swings of business. The curves of the index numbers for agriculture, industries manufacturing food products, and the stone, clay and glass manufacturing group showed non-typical fluctuations; price movements in these series did not correspond with the swings of general prices and of general business conditions.

The investigation was carried further, price movements of individual commodities being tested in the same way. All those price series which showed non-typical fluctuations, as well as those which were inflexible, were excluded as being unsuited to the purpose in mind. It was found that several series within non-typical industrial groups showed typical fluctuations; moreover, the prices of individual commodities were found to be more flexible and sensitive than group index numbers. For these reasons it was decided that the final price index of business cycles should be constructed from individual price series and not from group index numbers.

The commodities finally selected for inclusion were only ten in number. They included cotton-seed oil, coke, pig zinc, pig iron, bar iron, mess pork, hides, print cloths, sheetings and worsted yarns. The prices of these commodities are all highly flexible and sensitive to business

changes. The industries they represent, moreover, are important industrially.

When price quotations for these ten commodities had been secured extending back to 1890, the desired "price index of business cycles" was constructed. Prices for each month and year were expressed as relatives on the 1890–99 base. (The base has since been changed to 1919.) The index is an unweighted geometric mean of these relative prices.

Other Wholesale Price Index Numbers

Somewhat similar in character, though not constructed with the avowed purpose of serving as an index of business cycles, is the wholesale price index of 20 basic commodities compiled by the Federal Reserve Bank of New York. It was constructed to serve as a sensitive index of price changes in the basic commodity markets. As published in the *Monthly Review* of the Federal Reserve Bank of New York this index appears as a relative, with 1913 as the base. It has been computed only for the period 1913 to date.

An index of wholesale food prices is published weekly by the *Annalist*. It is based upon price quotations of 25 basic articles of food. These are reduced to relatives on the base 1890–99, and combined in the form of an unweighted arithmetic mean. The index purports to serve only as an index of changes in the cost, at wholesale, of the food called for by a theoretical family food budget.

Another index of wholesale food prices is published by Thomas Gibson of New York, and appears in his weekly market letter. It is based upon the prices of 22 primary articles of food. No full account of the weighting system has been given, though it is stated that Dun's general method has been employed. The index, as published, is based on weighted relative prices, the base of each being the average price during the period 1890–99. The result is modified, however, through multiplication by a constant factor, a

process originally designed to bring this index into harmony with Dun's, of which it was to serve as a continuation.

Differences Between Index Numbers

If the index numbers of wholesale prices which have been described are compared over a period of years [1] there is found to be a rather surprising agreement as to the general direction of price movements, but certain significant differences in details appear. Thus one index number may show an increase in the price level from one month to the next, while another index may show no change, or even a decline. Such differences are accentuated in periods of violent price fluctuations, such as those which occurred from 1916 to 1921. The reasons for such differences have been suggested in the discussion of methods of making index numbers. In every case such contradictory movements may be traced to one or more of the following:

a. Differences in the number of commodities.
b. Differences in the kinds of commodities included.
c. Differences in the method of securing price quotations (e.g., contract or open market; first of month or average of monthly prices).
d. Differences in weights.
e. Differences in methods of averaging.

Summary: Index Numbers of Wholesale Prices

The following summary of the main characteristics and uses of the different index numbers is based upon a consideration of these differences.

1. Of indices currently published the best general purpose index number of wholesale prices in the United States is that of the United States Bureau of Labor Statistics. It is a measure,

[1] Such a comparison of the three chief American index numbers of wholesale prices is made by W. C. Mitchell in "The Making and Using of Index Numbers," Part I, *Bulletin 284*, U.S. Bureau of Labor Statistics, 94–112.

however, not of the average ratio of change in prices, but of variations in the money cost of a constant quantity of goods.

2. For the period 1913–1918 the index of the War Industries Board is the best general index of wholesale prices.

3. Dun's index is a good measure of wholesale price changes, but its value is lessened somewhat by the undue weight given to food products and by the failure of its makers to announce the commodities included and the exact weights employed.

4. Bradstreet's index is a good measure of wholesale price changes, but the weight given to raw materials detracts from its representative value. The absence of logical weights is another weakness of some importance. Because of the weight given to raw materials this index serves as a better business barometer than either of the other two major indices.

5. Persons' index is not designed to be a general measure of price changes. It is well adapted to perform the function for which it is constructed, that of serving as an index of business cycles.

6. The index of 20 commodities constructed by the Federal Reserve Bank of New York serves as a useful measure of changes in prices of basic materials.

7. The indices of the Federal Reserve Board, constructed for the purpose of international comparison, perform a highly useful function in facilitating comparison of price movements in different countries. The separate index numbers constitute good measures of general price movements in the various countries for which they are constructed.

8. Fisher's weekly index of wholesale commodity prices serves a very useful purpose as a comprehensive and accurate price index. Its utility is enhanced by its prompt publication.

9. The *Annalist* index of wholesale food prices is of limited value. The method of construction (securing an unweighted average of relative prices) is likely to give results of doubtful value, and the base period (1890–99) is too far removed.

Other Price Index Numbers

The measurement of price changes by the use of index numbers has not been confined to wholesale prices. Many

variations of this device have been utilized in measuring price movements in other fields. It will be useful at this point briefly to indicate the character of some of these variations.

Index Numbers of Retail Prices

An index of retail food prices is published currently by the United States Bureau of Labor Statistics. The general methods employed are similar to those already explained in connection with the index of wholesale prices computed by that agency, with such differences as inevitably result from the nature of the material.

On the 15th day of each month actual retail selling prices of 43 articles of food [1] are secured from dealers in 51 representative cities throughout the United States. Though there are variations in local practice comparable quotations are secured, in so far as this is possible. The total of the money prices for each article is taken, and the average for the United States secured by dividing this sum by the number of dealers reporting prices on that article. A weighted aggregate of these individual prices is then secured. Each article is weighted in accordance with its relative importance in the consumption of the average family in 1918, as determined by various budget studies conducted by the Bureau. The method of securing weighted aggregates of actual prices resembles that employed in the construction of the wholesale price index. The retail price aggregates are also turned into relatives on the 1913 base, and these relatives are currently published in the *Monthly Labor Review*.

Similar operations are carried through for the 51 individual cities, and an index number of retail food prices is published for each of these cities. The actual commodity prices, averaged for the United States and for each of the

[1] The index has been based upon 43 articles of food since January, 1921; prior to that date the index was based upon 22 articles of food.

cities, are published in occasional retail price bulletins, a practice which adds materially to the value of the work done by the Bureau of Labor Statistics.

The difficulties inherent in the problem of measuring wholesale price movements have been discussed at some length. The construction of index numbers of retail prices of the type just described presents even greater difficulties. All the theoretical problems arising in the former case are to be solved and, in addition, the practical difficulties of securing suitable weights, accurate price figures and comparable quotations are intensified. Because of the lack of commodity standardization, and because of variations in business practice and local customs, the latter difficulty is particularly acute. For these reasons no index of retail prices at present published can be accepted with the confidence with which the best indices of wholesale prices may be received.

Index Numbers of the Cost of Living

If these problems are acute in constructing an index of retail prices they are doubly hard to solve in measuring such an entity as the cost of living. When food prices, rents, retail clothing prices, cost of fuel and light, retail furniture prices and the prices of the other miscellaneous items which are included in the budget of the average family are to be averaged, and an index number constructed to measure variations in the cost of these items, numerous statistical difficulties must be overcome. Theoretical questions concerning the most suitable methods of averaging and weighting present themselves, but more important are the practical problems involved in the collection of accurate and comprehensive data. Index numbers of the cost of living, therefore, must be used with particular caution until their reliability has been more effectively demonstrated than has been the case to date.

Two index numbers of the cost of living are currently compiled, one by the Bureau of Labor Statistics, one by the National Industrial Conference Board, an organization of employers' associations. The former appears in the *Monthly Labor Review*, the latter in publications of the National Industrial Conference Board. In each case the chief items of domestic expenditure are weighted in accordance with their relative importance in household budgets, and the combined result expressed as a relative, with 1913 values as the base. (July, 1914, is the base for the Industrial Conference Board index.)

Index Numbers of Price and Buying Power of Farm Products

A set of very useful index numbers is compiled by the United States Department of Agriculture. One of these is designed to measure, by months, changes in the prices received at the farm for the chief agricultural crops. Another measures changes in the farm prices of live-stock products. The two are combined into a third index measuring changes in the average price of all farm products. All these index numbers are published as relatives, with 1913 as base.

The significance of changes in the index of prices of farm products is made clearer by the computation of a fourth index of purchasing power of farm products. The index of money prices means little by itself, since the purchasing power of money is constantly changing. Accordingly, it must be interpreted in terms of an index which measures changes in the prices of the things the farmer buys with his money. Such an index is secured by re-compiling the index of wholesale prices of the Bureau of Labor Statistics, excluding farm products and foods. The computation of the index of purchasing power of farm products may be illustrated, with reference to the figures for January, 1923.

For that month the index of prices of farm products was 116 (1913 = 100). The index of commodity prices at wholesale, excluding farm products and foods, was 170. That is, the farmer was receiving 16 per cent more for his products than in 1913, but general commodities at wholesale cost 70 per cent more than in 1913. Therefore his purchasing power, with reference to 1913 as base, was equal to $\frac{116}{170} \times 100$, or 68.2 per cent.

The values of all these index numbers, by months, for the year 1923, are given in the following table.

TABLE 57

Commodity Price Indices

Index Numbers of Price and Buying Power of Farm Products, 1923
(1913 = 100)

Month	*Prices at the Farm*			* *Wholesale Price of Commodities*	† *Purchasing Power of Farm Products*
	Crops, 15th of Month	*Live-Stock, 15th of Month*	*Crops and Live-Stock Combined*		
January	126	106	116	170	68
February	130	107	118	172	69
March	134	106	120	175	69
April	139	107	123	176	70
May	140	105	123	172	71
June	139	100	120	168	71
July	136	102	119	165	72
August	136	102	119	163	73
September	138	109	123	164	75
October	139	103	121	161	75
November	137	97	117	160	73
December	137	94	116	158	73

* Excluding farm products and foods. † Expressed in terms of other products.

It should be clear that one important assumption is made in the computation of the above index of purchasing power. This is, that changes in the prices of the things the farmer buys are measured by the corrected index of wholesale

commodity prices of the Bureau of Labor Statistics. This may be questioned, for the farmer does not generally buy at wholesale, nor does he buy just the things included in that index. It is used, of course, in default of a more appropriate index, but the possible errors it introduces should be recognized.

These index numbers have been compiled for the period from 1913 to date, by months. They are published in *Crops and Markets*, issued by the Department of Agriculture.

Index Numbers of Money Wages and Real Wages

In years when industrial troubles are numerous particular interest attaches to the course of wages. And though the interest is intensified at these times, the problem is one of permanent importance. Many index numbers of wages have been prepared by governmental agencies, employers' associations, labor unions, and private investigators. No attempt to evaluate these various indices will be made at the present time; for our present purpose it is sufficient that a few points of importance concerning the construction and interpretation of such index numbers be stressed.

The problem involved in the construction of an index of money wages is essentially one of averaging prices, the prices in this case being those of labor instead of commodities. Difficulties at once arise because of the various ways in which the "prices" of labor are quoted. Shall the index of wage movements be based upon hourly rates of pay, full time weekly earnings (i.e., the amount which would be earned at prevailing hourly rates if a full time week were put in), or actual weekly earnings? An index based upon hourly rates would take no account of changes in the hours of labor, unemployment, under-employment, overtime, or special bonuses. Changes in the hours of labor would be reflected in an index of the second type, but the other

objections would still hold. An index of actual weekly earnings, if it covered seasons of slack as well as full work, would appear to be the most useful type. But data for this sort of index are exceptionally hard to obtain, a fact of considerable practical importance. Such an index is constructed by the New York State Department of Labor, figures being secured from representative factories throughout the State. This index is given as an actual average in dollars and cents, though these figures may be reduced to relatives on any base chosen.

Apart from the type of wage figure which shall be used, there are questions as to methods of weighting and averaging which need not be discussed here. Probably of greater importance is the matter of "representativeness," corresponding to the problem as to the kinds of commodities to be included in a wholesale price index. If wages of skilled workmen alone are averaged the result cannot be presented as an index of general wage rates. Again, if the wages of foremen and highly paid technical workers are included in a general wage index they should be given only the weight they deserve and the nature of the resulting index should be clearly explained. In fact, in this case, as with wholesale prices, it is highly desirable that separate index numbers be constructed for the different classes of wage-earners, these being later combined into a properly weighted average of all wages. Our knowledge of the nature of wage movements, as of price movements, is more likely to be extended by breaking up the materials to be averaged into significant and homogeneous groups and studying these groups in detail, than by perfecting a single index of a great number of heterogeneous classes.

Differences in practice in regard to some of the points suggested above have led to numerous and important differences between published index numbers of wages. Unless the methods employed in the construction of such index numbers are fully explained and the nature of the

material entering into them completely described, the story of wage movements which they tell may only be accepted with reservations.

An index of money wages would be significant in itself only if the cost of living remained unchanged. But when prices and the cost of living are rising, the gains registered by a rising index of wages may be purely illusory; the same is true of losses under opposite conditions. If a measure of relative well-being is desired it is necessary to convert the index of money wages into an index of real wages, which shall take account of changes in the cost of living.

The example which follows will serve to illustrate the process of constructing an index of real wages. The index of wages is based upon average weekly earnings of approximately 475,000 workers in New York State, earnings for December, 1914, being taken as the base. The cost of living index is that of the United States Bureau of Labor Statistics for thirty-two cities throughout the country, the figure for December, 1914, being taken as 100.[1]

TABLE 58

Index Numbers of Real Wages in New York State, 1914–1923

(December, 1914 = 100.)

Date	*Relative weekly earnings of New York workers*	*Cost of living*	*Real Wages (i.e., Purchasing power of money wages.)*
December, 1914	100	100	100
" 1915	107	102	105
" 1916	123	115	107
" 1917	140	138	101
" 1918	185	169	109
" 1919	209	193	108
" 1920	225	195	115
" 1921	198	169	117
" 1922	210	165	127
" 1923	222	168	132

[1] It would be more logical in the present instance to correct the New York wage figures by a cost of living index for New York cities alone. The difference in the results would not be material, however.

In securing the index of real wages, the index of relative earnings is divided by the cost of living index for the corresponding date. It is clear that the earnings figure alone is quite inaccurate as a measure of the condition of the workers. When corrected it shows an increase in real wages, but one much less pronounced than the gain in money wages. Such figures at their best, however, may be accepted as only approximations to the truth. There is a considerable margin of error in the construction of both sets of index numbers from which the measure of real wages is computed, so that the latter should only be looked upon as a rough index of changes in the real income of New York factory workers.

REFERENCES

Bowley, A. L. *Elements of Statistics* (196–213).

Chaddock, R. E. *Principles and Methods of Statistics* (Chap. X).

Fisher, Irving. *Revision of the Weekly Index Number.* Journal of the American Statistical Association, Sept., 1924.

Fisher, Irving. *The Making of Index Numbers.*

Flux, A. W. *The Measurement of Price Changes.* Journal of the Royal Statistical Society, March, 1921 (167–215).

Kelley, Truman L. *Statistical Method* (331–347).

Mitchell, W. C. *The Making and Using of Index Numbers.* Part I, Bulletin 284, U. S. Bureau of Labor Statistics. Oct., 1921.

Persons, Warren M., and Coyle, Eunice. *A Commodity Price Index of Business Cycles.* Review of Economic Statistics. Prel. Vol. III (353–369).

Persons, Warren M. *Fisher's Formula for Index Numbers.* Review of Economic Statistics. Prel. Vol. III (103–113).

Walsh, C. M. *The Measurement of General Exchange Value.*

Walsh, C. M. *The Problem of Estimation.*

Young, Allyn A. *The Measurement of Changes in the General Price Level.* Quarterly Journal of Economics, Aug., 1921 (557–573).

Young, Allyn A. *Index Numbers* (in Rietz, H. L., *Handbook of Mathematical Statistics*, 181–194).

CHAPTER VII

THE ANALYSIS OF TIME SERIES: MEASUREMENT OF TREND

The preceding sections have dealt primarily with frequency series and with problems arising in the attempt to organize and describe such series. We are now concerned with data in the study of which the essential problem is the analysis of chronological variations. Such series are of major importance in the field of economic statistics, for most of the data of economics and business are variables in time — as bank clearings, steel production, volume of sales, etc. This dominating importance of series in time is not found in any other field of statistical research, and the development of methods of analysis appropriate to time series has come, accordingly, only within recent years with the wider adoption of statistical methods in the field of economics.

Problems connected with time series arise both in the ordinary routine of internal administration and in the analysis of general economic conditions. Sales, purchases, profits on the one hand, stock prices, interest rates, business failures on the other, are variables which fluctuate with the passage of time. In the analysis of such series it is generally desired that the rate and character of growth be determined, and that periodic and accidental fluctuations be isolated for study. The sales manager wishes to know how the volume of sales is faring, when and why it fluctuates and how it compares with volume of production. The economist desires to know the trend of prices, and to scrutinize minutely the upward and downward movements of the price level. The making of business plans on even a small scale,

as well as the most elaborate schemes of economic forecasting, must rest upon such study of past trends and fluctuations, and upon comparison of the movements of related series in time. Scientific study of the business cycle is only possible through the application of such methods. Our present task is the development of methods appropriate to the analysis of series in time.

The Preliminary Organization of Time Series

The data of time series usually require less preliminary organization than statistical data which are to be reduced to the form of a frequency distribution. The source, primary or secondary, from which the figures are taken usually presents them in shape for analysis. Certain precautions should be observed, however.

The dates to which the figures apply should be clearly understood and definitely stated. Monthly data may be as of the first of each month (as in the case of Bradstreet's index number of wholesale prices), averages for each month (as in the case of the Bureau of Labor Statistics' price index), or totals for each month (as in the case of figures for cotton consumption). They may be cumulative monthly figures, each item representing the total for the year to date, as in the case of certain coal production data. If average figures are given for a month or year it is important to know how the average has been secured.

Again, it is essential that in any time series there be strict comparability between data for different dates. Any attempt to analyze a series which is not homogeneous must be misleading and futile. Yet such series are not infrequently published. Commodity production or consumption figures published by trade associations and by governmental agencies are often based upon returns from a varying number of reporting concerns. A series of price quotations may lack comparability as between different dates because

of slight changes in the unit or grade to which the quotation applies, or because quotations are drawn from different markets. Changes in census classifications may result in lack of comparability in census data. Merchandise import and export figures are reported to have been compiled, until recently, in ways which render misleading comparison between the figures for different periods. A change in a salesman's territory may alter his returns materially. It is stated that the character of the obligations represented by the United States Steel Corporation's figures for "unfilled tonnage" varies from time to time. Thus in July, 1922, it was reported, the figures meant actual orders for shipment at mill convenience, while a year earlier much of the business on its books was in contract form, with date of shipment not specified. These are examples of faults which may be found in time series, rendering analysis futile. Strict testing is essential before a series be accepted as accurate and homogeneous.

Graphic Representation of Time Series

Normally the first step to be taken in visualizing a series in time and in preparing for further analysis consists of plotting the data. The trend and general characteristics of a series may be most readily apprehended through graphic representation. The data may be plotted on ordinary arithmetic paper, on semi-logarithmic paper or, more rarely, on double logarithmic paper. The advantages of the latter types for certain purposes have already been explained. The choice in a given case will depend upon the nature of the data and the object of the study. If interest lies in the absolute amount of fluctuations in sales, prices, pig iron production or whatever may be in process of analysis, or in the comparison of absolute differences between series, the ordinary rectilinear chart is to be employed. If percentage variations and the comparison of relative

fluctuations are matters of interest the semi-logarithmic representation is preferable. In general, if one is accustomed to the interpretation of this latter type of chart, its use is advisable. A clearer, less-distorted presentation of relations and a more significant comparison of series are generally secured when economic data having time as one variable are plotted on the semi-logarithmic paper.

For some purposes the process of studying series in time will have been completed when the data are thus plotted. The general trend may be roughly determined from the chart. The existence of seasonal and other periodic variations may be ascertained. Rough comparisons of trends and fluctuations may be made. All the knowledge thus secured, it should be noted, will be non-quantitative in character, and the comparisons will be tentative and approximate. Even so, such charts enable trends and relations to be much more clearly visualized than do the raw figures, and for some purposes the knowledge thus secured is sufficient, though it lacks precision and accuracy. For other purposes more exact measurement and more refined analysis are required. Certain appropriate methods may be described.

Forces Affecting Series in Time

The general object in the analysis of a time series is the isolation of the effects of one or more of the forces affecting the given series. This may be desired in order that the past behavior of the single series may be understood, in order that the future behavior of the series may be predicted, or in order that two or more series may be compared. It is not in any case possible to isolate these effects of individual forces with absolute accuracy, and in some cases it is impossible even to approximate such a result. But, in general, given figures covering a sufficiently long period, the effects of various influences upon the behavior of a given

series may be measured with some degree of accuracy. The possibilities of using such measurements in economic analysis have by no means been fully realized.

What are these forces which affect series of data in time? The forces in any given case may be unique, affecting only the given series, but in general the various influences acting upon such series may be placed in a limited number of categories.

Secular Trend

In the first place, most series of economic statistics exhibit a definite *trend*, which may be constant in direction, changing direction at a constant rate, or even characterized by abrupt changes in direction or rate which are due to the introduction of novel elements. Thus the volume of production or sales of a business house, taken over a period of years, will usually show a fairly regular growth. The same is true of population, the production of basic minerals, the number of motor vehicles registered, etc. In some cases the rate of growth may be a negative one, as in the case of interest rates in the United States over the last half century. The concept of *secular trend* (i.e., trend over a long period of time) covers both positive and negative changes of this type.

In the analysis of a time series the trend value at any date is taken to be the *normal* value at that date. That is, it is viewed as the value which would be recorded for the given series if the effects of all accidental and complicating forces could be eliminated, leaving only the effect of normal growth. This conception of a normal value for any series at a given date, a value which may be used as a base or point of reference in judging the effects of all forces other than the growth factor, is fundamental in economic analysis. "No other method," says Carl Snyder, "enables us so quickly to set economic events in their just perspective."

The fact should be emphasized that by secular trend is meant the smooth, regular, long-term movement of a statistical series. Frequent and sudden changes either in absolute amounts or in rates of increase or decrease are quite inconsistent with the concept of secular trend. It is true that there may be occasional changes due to the interjection of a new element or the withdrawal of an old factor. But the breaking up into numerous sections of the period covered by a time series, and the determination of trend for each of these minor periods, does violence to the whole concept of gradual change over a period of time. If the factors affecting the series are subject to violent and frequent alteration, no secular trend can be assumed.

It does not follow from this discussion that a definite upward or downward trend exists for all time series. Many series, such as barometric readings at a certain point, merely fluctuate about a constant level which does not change with the passage of time.

Periodic Fluctuations

If the plotted representation of a time series be studied, the long-term trend may be discerned in the general upward or downward drift, but may not be precisely determined by inspection because of the existence of numerous fluctuations, superimposed upon the trend. These fluctuations may be regular or irregular, violent or mild, simple or complex. The value of the variable at any given date represents the net resultant of the interaction of the secular trend and the various forces which tend to pull the variable above or below its trend value. These latter forces, constituting the disturbing factors in a situation which would otherwise be characterized by gradual and normal change, may be of several types.

Seasonal variations are found in most series of economic statistics for which monthly values are obtainable. Con-

sumption and production of commodities, interest rates, bank clearings, railroad freight traffic and many other types of data are marked by seasonal swings repeated with minor variations year after year. These, in so far as they exist at all, are definitely periodic in character, with a constant twelve-month period. Less markedly periodic, but nevertheless characterized by a considerable degree of regularity, are the *cyclical fluctuations* which are found in series affected by forces connected with economic or business cycles. Prices, wages, the volume of industrial production, trading on the Stock Exchange, and most series relating to the activities of individual business units are affected by the swings of business through alternating periods of depression and prosperity. While the length of such periods may vary, the general sequence of change has been in the past sufficiently regular to render such movements capable of study.

Random Fluctuations

Entangled with these more or less regular movements are the effects of random, accidental and irregular fluctuations—catastrophic events such as the San Francisco earthquake, wars, floods, fires and countless minor events equally fortuitous though less violent in the resulting disruptions. These, like the seasonal or cyclical fluctuations, may serve to pull the actual value of a variable at a given date away from that normal value which would be expected were the regular and uniform change due to secular movement alone taking place. The net resultant of all these disturbing values determines how far above or below normal the actual value on a given date may be.

The analysis of series in time involves the isolation of the effects of these various forces, in so far as this is possible. A problem may call for the study of but one factor, or it may require the complete breaking up of given values. When annual data are used the seasonal element will not

enter, of course. The explanation of methods begins with a consideration of problems involving only this type of data.

The Measurement of Secular Trend

As an example of the type of material in connection with which these problems arise, the following figures may be taken.[1] The values are given in thousands of millions in order to simplify the calculations.

Table 59

Bank Clearings in New York City, 1860–1923

(in thousands of millions)

1860 $ 7.2	1876 $21.6	1892 $ 36.7	1908 $ 79.3
1861 5.9	1877 23.3	1893 31.2	1909 103.6
1862 6.9	1878 19.9	1894 24.4	1910 97.3
1863 14.9	1879 29.2	1895 29.9	1911 92.4
1864 24.1	1880 38.6	1896 28.8	1912 100.7
1865 26.0	1881 49.4	1897 33.4	1913 94.6
1866 28.7	1882 46.9	1898 42.0	1914 83.0
1867 28.7	1883 37.4	1899 60.8	1915 110.6
1868 28.5	1884 31.0	1900 52.7	1916 159.6
1869 37.4	1885 28.2	1901 79.4	1917 177.4
1870 27.8	1886 33.7	1902 76.3	1918 178.6
1871 29.3	1887 33.4	1903 66.0	1919 235.8
1872 33.8	1888 31.1	1904 68.6	1920 243.2
1873 35.5	1889 35.9	1905 93.8	1921 194.4
1874 22.9	1890 37.4	1906 104.7	1922 217.9
1875 25.1	1891 33.7	1907 87.2	1923 214.0

As has been pointed out, the figure for any year, as the value of $100.7 thousands of millions for 1912, is the net resultant of many forces which may be classified as the secular trend, cyclical variations and random or accidental fluctuations. Our first problem is to determine the effect of the first of these forces.

In Fig. 54 the data of New York bank clearings during the period 1860–1923, inclusive, have been plotted. A definite trend is apparent, together with well marked and

[1] It is generally advisable to deflate such series as this before attempting further analysis. The procedure is explained below.

more or less regular deviations from that trend. Several methods are available for arriving at approximations to this trend. By employing moving averages an attempt may be made to eliminate the effect of passing fluctuations and to arrive at values which represent the effect of the steadily

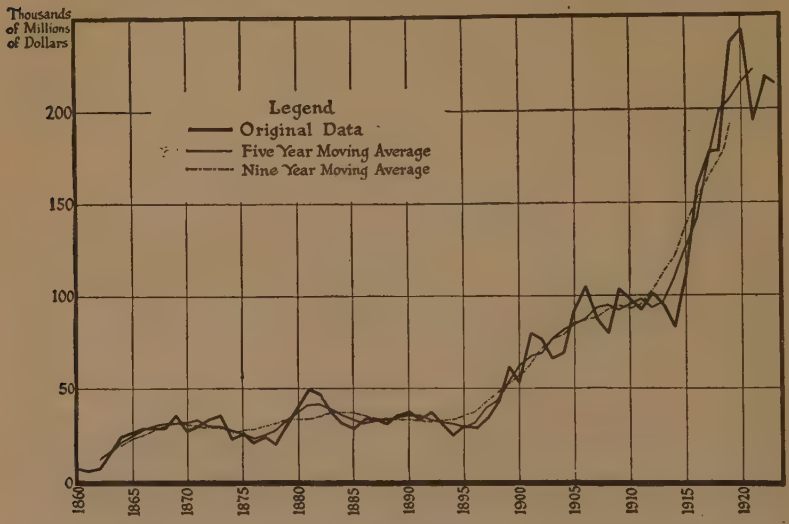


FIG. 54. — New York Bank Clearings, 1860–1923, with Moving Averages

operating growth factor. Assuming a definite functional relationship between the time factor and the other variable to prevail, empirically at least, an approximation to the trend may be secured by fitting an appropriate curve to the plotted data. Smoothing the data by hand gives somewhat the same result, the curve being frankly approximative and empirical in character. In certain studies it has been found possible to use one statistical series as base or trend line for another series of homogeneous data.

MOVING AVERAGES

When a trend is to be determined by the method of moving averages, the average value for a number of years (or months, or weeks) is secured, and this average is taken

as the normal or trend value for the middle of the period covered in the calculation of the average. The following table shows the results secured when three-, five-, seven-, and nine-year moving averages are thus computed for New York Bank clearings for the period 1900–1923:

TABLE 60

New York City Bank Clearings, 1900–1923, and 3-, 5-, 7-, and 9-year Moving Averages

(in thousands of millions)

Year	*Original data*	*Three-year moving av.*	*Five-year moving av.*	*Seven-year moving av.*	*Nine-year moving av.*
1900	$ 52.7				
1901	79.4	$ 69.5			
1902	76.3	73.9	$ 68.6		
1903	66.0	70.3	76.8	$ 77.4	
1904	68.6	76.1	81.9	82.3	$ 78.7
1905	93.8	89.0	84.1	82.3	84.3
1906	104.7	95.2	86.7	86.2	86.3
1907	87.2	90.4	93.7	90.6	88.1
1908	79.3	90.0	94.4	94.0	92.0
1909	103.6	93.4	92.0	95.0	94.8
1910	97.3	97.8	94.7	93.6	93.6
1911	92.4	96.8	97.7	93.0	94.3
1912	100.7	95.9	93.6	97.5	102.3
1913	94.6	92.8	96.3	105.5	113.2
1914	83.0	96.1	109.7	116.9	121.6
1915	110.6	117.7	125.0	129.2	137.0
1916	159.6	149.2	141.8	148.5	153.7
1917	177.4	171.9	172.4	169.7	164.1
1918	178.6	197.3	198.9	185.7	177.8
1919	235.8	219.2	205.9	201.0	192.4
1920	243.2	224.5	214.0	208.8	
1921	194.4	218.5	221.1		
1922	217.9	208.8			
1923	214.0				

The three-year moving average for 1904 is the average of the figures for 1903–4–5, the five-year figure is the average of the years 1902–3–4–5–6. The other averages are computed in the same way. In each case the average is centered for the period included; that is, the average is taken to

represent normal as of the middle of the given period. The employment of an odd number of years simplifies this centering process, though it is not essential that the number be odd. With an even number of years, the figure may be centered by taking a two-year moving average of the moving average. The five- and nine-year moving averages for the entire period are plotted, with the original data, in Fig. 54.

It is obvious that the effect of the averaging is to give a smoother curve, lessening the influence of the fluctuations which pull the annual figures away from the general trend. The longer the period included in securing each average, the smoother is the curve secured, though there are other factors to consider in deciding upon the length of the period. Certain of these factors may be noted.

Characteristics of Moving Averages

Given cyclical fluctuations about a uniform level or about a line ascending with a uniform slope, the length of the cycle and the magnitude of the fluctuations being constant, a moving average having a period equal to the period of the cycle (or to a multiple of that period) will give a straight line, a perfect representation of the trend. Under the same conditions a moving average having a period greater or less than the period of the cycle will give, not a straight line, but a new cycle having the same period as the original, but with fluctuations of less magnitude. The minima and maxima of the cycles thus obtained, moreover, will not necessarily coincide with the minima and maxima of the original cycles. In general, when such a new cycle is obtained the magnitude of the fluctuations will be less the longer the period on which the average is based.[1]

These propositions may be illustrated by the following

[1] The decrease in the magnitude of the fluctuations is not regular, however, but cyclical.

figures, arbitrarily chosen. In the first example five figures have been selected which repeat themselves in sequence, fluctuating about a common level.

TABLE 61

Illustrating the Application of Moving Averages

(1)	(2)	(3)	(4)	(5)
Cyclical Data	*Moving average of 5 items*	*Moving average of 10 items (centered)*	*Moving average of 3 items*	*Moving average of 8 items (centered)*
2				
6			$5\frac{1}{3}$	
8	$6\frac{1}{5}$		8	
10	$6\frac{1}{5}$		$7\frac{2}{3}$	
5	$6\frac{1}{5}$		$5\frac{2}{3}$	$6\frac{3}{8}$
2	$6\frac{1}{5}$	$6\frac{1}{5}$	$4\frac{1}{3}$	$6\frac{13}{16}$
6	$6\frac{1}{5}$	$6\frac{1}{5}$	$5\frac{1}{3}$	$6\frac{3}{8}$
8	$6\frac{1}{5}$	$6\frac{1}{5}$	8	$5\frac{6}{8}$
10	$6\frac{1}{5}$	$6\frac{1}{5}$	$7\frac{2}{3}$	$5\frac{11}{16}$
5	$6\frac{1}{5}$	$6\frac{1}{5}$	$5\frac{2}{3}$	$6\frac{3}{8}$
2	$6\frac{1}{5}$	$6\frac{1}{5}$	$4\frac{1}{3}$	$6\frac{13}{16}$
6	$6\frac{1}{5}$	$6\frac{1}{5}$	$5\frac{1}{3}$	$6\frac{3}{8}$
8	$6\frac{1}{5}$	$6\frac{1}{5}$	8	$5\frac{6}{8}$
10	$6\frac{1}{5}$	$6\frac{1}{5}$	$7\frac{2}{3}$	$5\frac{11}{16}$
5	$6\frac{1}{5}$	$6\frac{1}{5}$	$5\frac{2}{3}$	$6\frac{3}{8}$
2	$6\frac{1}{5}$		$4\frac{1}{3}$	$6\frac{13}{16}$
6	$6\frac{1}{5}$		$5\frac{1}{3}$	
8	$6\frac{1}{5}$		8	
10			$7\frac{2}{3}$	
5				

(The items in columns (3) and (5) have been centered by means of a moving average of 2 items.)

The moving averages in columns (2) and (3) represent the data with the cycles completely removed. When the period of the average is not equal to the period of the cycle, or to a multiple of that period, the cycle is not removed, as is apparent from the figures in columns (4) and (5).

The conclusions suggested above hold when the cyclical fluctuations take place about any straight line. In the example which follows the foregoing data have been employed but with a constant increment of 3. This is equiva-

lent to superimposing the same cycles upon a line with a slope of + 3.

TABLE 62

Illustrating the Application of Moving Averages to a Series with Linear Trend

(1)	(2)	(3)	(4)	(5)
Cyclical Data	*Moving average of 5 items*	*Moving average of 10 items (centered)*	*Moving average of 3 items*	*Moving average of 8 items (centered)*
2				
9			$8\frac{1}{3}$	
14	$12\frac{1}{5}$		14	
19	$15\frac{1}{5}$		$16\frac{2}{3}$	
17	$18\frac{1}{5}$		$17\frac{2}{3}$	$18\frac{3}{8}$
17	$21\frac{1}{5}$	$21\frac{1}{5}$	$19\frac{1}{3}$	$21\frac{1}{16}$
24	$24\frac{1}{5}$	$24\frac{1}{5}$	$23\frac{1}{3}$	$24\frac{3}{8}$
29	$27\frac{1}{5}$	$27\frac{1}{5}$	29	$26\frac{6}{8}$
34	$30\frac{1}{5}$	$30\frac{1}{5}$	$31\frac{2}{3}$	$29\frac{1}{16}$
32	$33\frac{1}{5}$	$33\frac{1}{5}$	$32\frac{2}{3}$	$33\frac{3}{8}$
32	$36\frac{1}{5}$	$36\frac{1}{5}$	$34\frac{1}{3}$	$36\frac{1}{16}$
39	$39\frac{1}{5}$	$39\frac{1}{5}$	$38\frac{1}{3}$	$39\frac{3}{8}$
44	$42\frac{1}{5}$	$42\frac{1}{5}$	44	$41\frac{6}{8}$
49	$45\frac{1}{5}$	$45\frac{1}{5}$	$46\frac{2}{3}$	$44\frac{1}{16}$
47	$48\frac{1}{5}$	$48\frac{1}{5}$	$47\frac{2}{3}$	$48\frac{3}{8}$
47	$51\frac{1}{5}$		$49\frac{1}{3}$	$51\frac{1}{16}$
54	$54\frac{1}{5}$		$53\frac{1}{3}$	
59	$57\frac{1}{5}$		59	
64			$61\frac{2}{3}$	
62				

(The items in columns (3) and (5) have been centered by means of a moving average of 2 items.)

The trend values, with the effect of the cycles completely removed, are secured by taking moving averages equal in period to the cycle or to a multiple of that period. The cycle persists, with the same period but with diminished amplitude, when the average is based upon a period not equal to that of the cycle, as is clear from the figures in columns (4) and (5).

When these ideally simple conditions of constant period and amplitude do not exist, the moving average becomes

more ambiguous and its interpretation becomes less simple. If the period of the cycle varies, the selection of a period for the moving average is more difficult. In general, a period equal to or greater than the average length of the cycle is to be selected. An average having a shorter period will give a line which is marked by pronounced cycles, these cycles being reduced as the period covered in the calculation of the average increases.

When the amplitude of the cycle varies, the period being constant, a moving average with a period equal to the length of the cycle will give a line of trend marked by minor cycles. The amplitude of these secondary cycles will be a minimum when the period of the average is equal to the period of the cycle (or to a multiple of that period). When these last two irregularities are combined, and the data are characterized by cycles of varying amplitude and of varying length, the moving average giving the most effective representation of the trend is that which has a period equal to the average length of the cycle, or to a multiple of that length.

A new factor enters when the trend departs from linearity. If the underlying trend of a series is concave upward, a moving average will always exceed the actual trend value; if the reverse is true, and the trend is convex upward, a moving average will always be less than the actual trend value.

These conditions are depicted in the following examples. In Table 63 figures are presented which give the values secured when a cycle of constant period and amplitude is superimposed upon a line of trend which is concave upward, i.e., increasing at a constantly increasing rate. If the moving average could completely eliminate the effects of the cycle, the values secured from the average would be equal to the average value of the five items ($6\frac{1}{5}$) plus the values of the function $y = x^2$.

TABLE 63

Illustrating the Application of Moving Averages to a Non-Linear Series

(Increasing rate)

(1)	(2)	(3)	(4)	(5)	(6)
x	x^2	*Cyclical Data*	*Col. (2) plus col. (3)*	*Moving average of 5 items*	*True trend values* $(x^2+6.2)$
0	0	2	2		
1	1	6	7		
2	4	8	12	12.2	10.2
3	9	10	19	17.2	15.2
4	16	5	21	24.2	22.2
5	25	2	27	33.2	31.2
6	36	6	42	44.2	42.2
7	49	8	57	57.2	55.2
8	64	10	74	72.2	70.2
9	81	5	86	89.2	87.2
10	100	2	102	108.2	106.2
11	121	6	127	129.2	127.2
12	144	8	152	152.2	150.2
13	169	10	179	177.2	175.2
14	196	5	201	204.2	202.2
15	225	2	227	233.2	231.2
16	256	6	262	264.2	262.2
17	289	8	297	297.2	295.2
18	324	10	334		
19	361	5	366		

The values of the moving average are, in this case, always above the true trend values, a form of distortion that will always occur with a series of this type.

In Table 64 are shown the results of superimposing the same cyclical values upon a line of trend which is convex upward, i.e., increasing at a constantly decreasing rate. In this case, a perfect method of eliminating the cycles would give results equal to the average value of the five items ($6\frac{1}{5}$) plus the values of the function $y = \sqrt{x}$.

In this case the moving average values are consistently too low. The discrepancy is most marked for the lower values of x, as the decrease in the rate of growth is most marked for these values.

TABLE 64

Illustrating the Application of Moving Averages to a Non-Linear Series

(Decreasing rate)

(1)	(2)	(3)	(4)	(5)	(6)
x	$\sqrt{x}$	*Cyclical Data*	*Col. (2) plus col. (3)*	*Moving average of 5 items*	*True trend values* $(\sqrt{x}+6.2)$
0	0	2	2.00		
1	1.00	6	7.00		
2	1.41	8	9.41	7.428	7.61
3	1.73	10	11.73	7.876	7.93
4	2.00	5	7.00	8.166	8.20
5	2.24	2	4.24	8.414	8.44
6	2.45	6	8.45	8.634	8.65
7	2.65	8	10.65	8.834	8.85
8	2.83	10	12.83	9.018	9.03
9	3.00	5	8.00	9.192	9.20
10	3.16	2	5.16	9.354	9.36
11	3.32	6	9.32	9.510	9.52
12	3.46	8	11.46	9.658	9.66
13	3.61	10	13.61	9.800	9.81
14	3.74	5	8.74	9.936	9.94
15	3.87	2	5.87	10.068	10.07
16	4.00	6	10.00	10.194	10.20
17	4.12	8	12.12	10.318	10.32
18	4.24	10	14.24		
19	4.36	5	9.36		

Considerations previously reviewed have indicated that a moving average should, in general, be based upon a period at least equal to the period of the cycle, and preferably equal to some higher multiple of that period when the data are at all irregular. The longer the period covered, the greater the stability of the average. But when the underlying trend departs materially from the linear form, following a curve bending upward or downward, the error involved in the use of any moving average increases as the period of the average increases. If a moving average is used in such a case to measure the trend, the period of the average should be the shortest which will serve to average out the cycles; equal, that is, to the average length of one cycle.

In practice, however, these various conditions are found in complicated combinations. The fact that cycles vary in amplitude and length calls for a moving average based upon a fairly long period. The fact that the trend of the data is usually non-linear calls for a short period average to lessen the upward or downward distortion. A consideration of some importance in practical work is that a moving average can never be brought up to date. The lag is less, of course, the shorter the period covered by the average. The selection of a period in a given case must rest upon a study of the actual data with these various considerations in mind.

It has been assumed in the preceding discussion that the purpose of the moving average is the representation of secular trend. The moving average may be used, also, in smoothing data for the purpose of eliminating random fluctuations. For this purpose a moving average based upon a period less than the average period of the cycle should be selected.

We may return now to the problem relating to New York Bank Clearings. A study of the lines marked out by the different moving averages in Fig. 54 reveals significant differences between them. The five-year average follows the graph of the original data most closely, as would be expected. The nine-year average marks out the smoothest line of trend, but, on the other hand, departs most widely from the data. This is particularly noticeable during the period from 1893 to 1898 and from 1911 to 1915, a condition due to the pronounced changes in the rate of growth of the series during these two periods. Except for these distortions the general trend seems to be most accurately represented by the nine-year average.

In determining the relative merits of the different moving averages we are aided by a knowledge of the course of business during the period covered. The volume of New York bank clearings is a sensitive index of general business

conditions, responding immediately to changes in speculative and industrial activity. Major and minor business cycles are reflected in this series. Knowing the number of cycles through which business has passed during the last half century (1870–1920), we may determine which of the moving averages serves best as a standard from which to measure cyclical deviations. In this case we are practically working backward from a known result, a method not always available.

If we take as a starting point in each cycle the year before business attains a normal condition after depression, the following cycles in general business activity may be distinguished: [1]

1870–1878	1904–1908
1879–1885	1908–1911
1886–1897	1911–1914
1898–1904	1915–

The cycles marked out by the three-year moving average are too numerous to enumerate. In fact, the deviations from this average are primarily accidental and minor fluctuations and should not be classed as cycles. Deviations from the seven- and nine-year averages mark out the following cycles:

Cycles of deviations from seven-year averages	*Cycles of deviations from nine-year averages*
1871–1878	1871–1878
1879–1884	1879–1887
1885–1888	1888–1897
1888–1897	1898–1900
1898–1900	1900–1903
1900–1903	1904–1907
1904–1907	1908–1914
1908–1911	1915–
1911–1914	
1915–1918	

[1] These cycles are based upon an index computed by W. F. Ogburn and Dorothy Thomas; *cf. Quarterly Publications of the American Statistical Association*, Sept., 1922, 324–340.

Certain of the differences between the cycles thus determined are undoubtedly due to differences in the materials analyzed. Some of the other differences are worthy of note. Deviations from the seven-year average show three more cycles than either of the other series. The cycles from 1885 to 1888 and from 1915 to 1918 are duplicated in neither of the other series, appearing as minor fluctuations. The cycle from 1898 to 1900 appears also in the deviations from the nine-year average, but not in the index used as the basis of comparison. The cycle from 1911 to 1914 is duplicated in this index, but not in the deviations from the nine-year average. The great increase during the war years distorted the nine-year average in such a way as to smooth out a real cycle in the data.

In summary, it may be said that, for the present data, a moving average with a period of less than seven years will not serve as a measure of trend. The nine-year average serves effectively, except for the latter part of the period when the rate of increase is changing sharply. The seven-year period appears to be somewhat short, emphasizing some fluctuations which are not of major importance.

In general, the moving average has the prime advantage of flexibility. The representation of secular trend by mathematical curves frequently involves the breaking up of a period into two or three subdivisions, and the fitting of separate curves to each. This necessarily results from changing conditions and sharply changing rates of growth or decline. Where such changes occur the moving average has the merit of flexible adaptation to the new conditions and is often a more effective measure of secular trend than curves fitted with great labor.

The simple moving average is sometimes modified by giving varying weights to the constituent items, weighting more heavily the items near the centers of the groups successively set up in computing the averages. Such a weighted moving average is called a *progressive mean*. The coeffi-

cients of terms in the binomial expansion are used as weights. In the preceding examples, which called for a moving average of five items, the weights employed would be 1, 4, 6, 4, 1. While the progressive mean has its uses in some types of analysis, it does not seem appropriate as a means of eliminating periodicities in time series, or of smoothing out random fluctuations in such series.

Representation of Secular Trend by Mathematical Curves

For many types of data the secular trend may be represented by a mathematical curve rather than by a line based upon a moving average. Thus, if the growth (or decline) is by constant absolute increments (or decrements) a straight line will serve as an exact representation of the trend. Or the growth may be by constant percentages, as in the case of capital increase, when a principal sum increases in accordance with the compound interest law. A curve of a definite mathematical form furnishes the best representation of this trend. In many series of economic statistics the data seem to conform to definite laws of growth, or decline, and where this is the case the task of analysis, interpretation and projection is materially assisted by securing a mathematical expression for the underlying law. In practically all cases, of course, there are departures from this law, deviations above and below the line of secular trend. These deviations, however, do not destroy the value of an equation which describes the underlying law of development.

There is one fundamental difference between the moving average as a measure of trend and such mathematical curves. The former implies no definite "law" to which the data are assumed to conform. It is based upon the data as given; if the general trend changes, the moving average follows the new trend. It is a flexible measure of trend, adapting itself

to changing conditions, purporting to be nothing more than an empirical approximation to the drift of the series. Mathematical curves fitted to economic series are, in fact, nothing more than empirical approximations also, but in a somewhat different sense. They assume a "law" of change underlying the variations, accidental and otherwise, which show upon the surface of the data. It is an empirical law which is assumed, it is true, but nevertheless there is postulated a uniform and consistent trend capable of mathematical expression. If such an assumption is to have any validity it is essential that the period during which the law is supposed to hold be homogeneous, that there be no material changes in the conditions affecting the series being studied. Thus an equation, let us say, is secured for the trend of gold production. If a radical change should take place in methods of extraction the trend of gold production would change materially and the former equation would no longer apply. Data covering the period before and after such a change would not be homogeneous, and a single equation for the trend during the whole period should not be secured.

In the practical approach to a problem involving the determination of secular trend the first task is the selection of the appropriate type of curve. This is perhaps the most difficult part of the work; certainly it is the part in which the element of personal judgment enters most directly. For there is no objective rule to follow, no fixed standard by which the most appropriate curve may be selected. Something more will be said on this subject after the characteristics of the chief types of curves and the methods of fitting them have been described. For the present it may be assumed that a curve similar to one of the types described in Chapter II, or of a related form, has been selected, and that we face the practical task of fitting it to the data.

Fitting a Straight Line; the Method of Least Squares

If the data, when plotted, show a trend which can best be represented by a straight line the task of fitting is merely the determination of the constants in an equation of the form $y = a + bx$. The values of a and b which will give a line following most closely the trend of the data are to be obtained. A simple illustration may serve to demonstrate the various methods which may be employed. Nine points (1, 3; 2, 4; 3, 6; 4, 5; 5, 10; 6, 9; 7, 10; 8, 12; 9, 11) are plotted in Fig. 55. Our problem is the fitting of a straight line to these points.

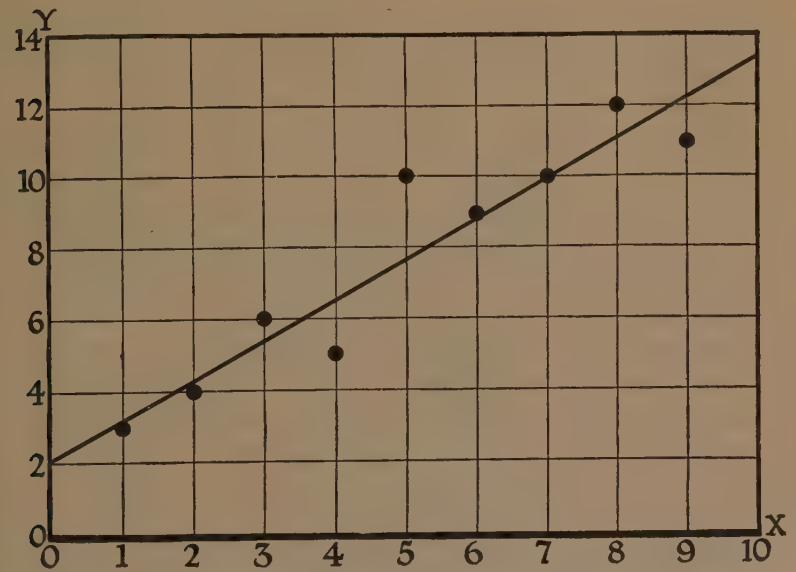


Fig. 55. — Illustrating the Fitting of a Straight Line to Nine Points

By inspection approximate values of a and b may be determined. A thread may be stretched through the points in such a direction that it seems to follow the trend as closely as possible. The slope of the line thus laid out

may be measured, the y-intercept determined, and the desired equation thus approximated. Obviously this is a loose and uncertain method, and the results obtained by different individuals may be expected to vary rather widely. There is one and only one straight line which fits most accurately the plotted data. The constants for this line of best fit may be determined by the method of least squares.

The theory upon which the method of least squares is based need not be detailed at length here. The argument may be briefly presented: A number of observation values of a certain quantity are found, and it is desired to obtain the most probable value of the quantity which is being measured. It is capable of demonstration that the most probable value of the quantity is that value for which the sum of the squares of the residuals is a minimum. (The "residual" is a term for the difference between a given value and each of the observation values.) This is true of the arithmetic mean of the observation values. Thus, if a given distance be measured by a number of individuals, with varying results, the most probable value is the arithmetic mean of the different measurements. The process of computing the mean involves the following steps, which are enumerated for the purpose of simplifying the later explanation. We seek a result, a statement of the most probable value of the distance being measured, which will take the form:

$$M = (\text{a constant}).$$

Let us say we have three approximations to this value:

$$M = 5672 \text{ feet}$$
$$M = 5671 \text{ feet}$$
$$M = 5676 \text{ feet}$$

adding, $3M = 17019$ feet

Since there is but one unknown, M, it may be derived directly from this equation, and we have

$$M = 5673 \text{ feet}.$$

This is the value for which the sum of the squares of the deviations is a minimum.

A similar problem arises when the relation between two variables is being measured. Our goal in this case is the equation which correctly describes this relationship. We have secured, however, varying results which do not agree precisely as to the constants in the equation of relationship. In other words, our plotted points do not all lie on the same line. What are the most probable values of the constants in the required equation? The answer is analogous to that given when a single quantity was being measured. We seek the constants which, when the resulting equation is plotted, will give a line from which the deviations of the separate points, when squared and totaled, will be a minimum. Assuming that each pair of measurements gives an approximation to the true relationship between the variables, we wish to find the most probable relationship, and this is given by the line for which the sum of the squared deviations is a minimum.[1]

We have, in the present example, nine pairs of values for x and y. Substituting these values in the generalized form of the linear equation, $y = a + bx$, we secure the following observation equations:

$$\begin{aligned} 3 &= a + 1b \\ 4 &= a + 2b \\ 6 &= a + 3b \\ 5 &= a + 4b \\ 10 &= a + 5b \\ 9 &= a + 6b \\ 10 &= a + 7b \\ 12 &= a + 8b \\ 11 &= a + 9b \end{aligned}$$

Any two of these equations could be solved as simultaneous equations, and values of a and b secured. But these values

[1] *Cf.* Appendix A for a more detailed discussion of the method of least squares, together with a description of certain checks upon the calculations.

would not satisfy the remaining equations. Our problem is to combine the nine observation equations so as to secure two *normal equations*, which, when solved simultaneously, will give the most probable values of a and b. The first of these normal equations is secured by multiplying each of the observation equations by the coefficient of the first unknown (a) in that equation, and adding the equations obtained in this way. Since the coefficient of a in the present case is 1 throughout, the nine observation equations are unchanged. The second of the normal equations is secured by multiplying each of the observation equations by the coefficient of the second unknown (b) in that equation, and adding the equations obtained. Thus the first equation is multiplied throughout by 1, the second by 2, and so on. The process of securing the two normal equations is illustrated below.

TABLE 65

Derivation of Normal Equations from Observation Equations

$3 = a + 1b$	$3 = 1a + 1b$
$4 = a + 2b$	$8 = 2a + 4b$
$6 = a + 3b$	$18 = 3a + 9b$
$5 = a + 4b$	$20 = 4a + 16b$
$10 = a + 5b$	$50 = 5a + 25b$
$9 = a + 6b$	$54 = 6a + 36b$
$10 = a + 7b$	$70 = 7a + 49b$
$12 = a + 8b$	$96 = 8a + 64b$
$11 = a + 9b$	$99 = 9a + 81b$
$70 = 9a + 45b$	$418 = 45a + 285b$

The two normal equations are

$$70 = 9a + 45b$$
$$418 = 45a + 285b.$$

It remains to solve these equations for a and b. By multiplying the first equation by 5 and subtracting it from the second, a may be eliminated; a value of $\frac{68}{60}$, or 1.133, is found for b. Substituting this value in either of the equa-

tions, a value of 2.111 is secured for a. The equation to the best fitting straight line is, therefore,

$$y = 2.111 + 1.133x$$

It is not necessary to write out and total the equations, as is done above. In the actual application of the method it is necessary only to insert the proper values in the two equations [1]

$$\Sigma(y) = na + b\Sigma(x)$$
$$\Sigma(xy) = a\Sigma(x) + b\Sigma(x^2).$$

The symbols employed have the following meanings:

$\Sigma(y)$: the sum of the values of y.
$\Sigma(x)$: the sum of the values of x.
$\Sigma(xy)$: the sum of the products of the paired x's and y's.
$\Sigma(x^2)$: the sum of the squares of the values of x.
n: the number of pairs of values; the number of points plotted.

The work of computation is facilitated by a tabular arrangement similar to the following:

TABLE 66

Computation of Values Required in Fitting a Straight Line

x	y	xy	x^2	
1	3	3	1	$n = 9$
2	4	8	4	$\Sigma(x) = 45$
3	6	18	9	$\Sigma(y) = 70$
4	5	20	16	$\Sigma(x^2) = 285$
5	10	50	25	$\Sigma(xy) = 418$
6	9	54	36	
7	10	70	49	
8	12	96	64	
9	11	99	81	
45	70	418	285	

The two desired normal equations are secured by substituting these five values in the type equations given

[1] General rules for the formation of normal equations are given in Appendix A.

above. It will be noted that the results are identical with those obtained from the observation equations.

When the equation to the best fitting straight line has been obtained the values of y corresponding to given values of x may be computed and compared with the observed values. The table which follows presents the results secured:

TABLE 67

Comparison of Observed and Computed Values of a Variable Quantity [1]

x	y (*observed*)	y (*computed*)	d	d^2	xd
1	3	$3.2\frac{4}{9}$	$-.2\frac{4}{9}$	.0597	$-.2\frac{4}{9}$
2	4	$4.3\frac{7}{9}$	$-.3\frac{7}{9}$	.1427	$-.7\frac{5}{9}$
3	6	$5.5\frac{1}{9}$	$+.4\frac{8}{9}$	.2390	$+1.4\frac{6}{9}$
4	5	$6.6\frac{4}{9}$	$-1.6\frac{4}{9}$	2.7041	$-6.5\frac{7}{9}$
5	10	$7.7\frac{7}{9}$	$+2.2\frac{2}{9}$	4.9381	$+11.1\frac{1}{9}$
6	9	$8.9\frac{1}{9}$	$+.0\frac{8}{9}$	.0079	$+.5\frac{3}{9}$
7	10	$10.0\frac{4}{9}$	$-.0\frac{4}{9}$	.0020	$-.3\frac{1}{9}$
8	12	$11.1\frac{7}{9}$	$+.8\frac{2}{9}$	.6760	$+6.5\frac{7}{9}$
9	11	$12.3\frac{1}{9}$	$-1.3\frac{1}{9}$	1.7190	-11.8
Total			0.0	10.4885	0.0

The sum of the deviations of the plotted points from the line is zero. The sum of the deviations when each is multiplied by the corresponding value of x is also zero. The accuracy of the actual calculations involved in fitting may be tested in this way. The sum of the squares of the deviations, 10.4885, is a minimum. Any change in the value of a or b would give a line from which the sum of the squared deviations would exceed 10.4885.

FITTING A STRAIGHT LINE; SPECIAL CASES

The simultaneous solution of the two normal equations will give, in any case, the most probable values of a and b. The processes of calculation may be simplified in certain special cases, not infrequently encountered in handling

[1] The common fractions are retained in certain columns in order that the sum of the deviations may be exactly zero.

economic data. If the x's are consecutive numbers, as they always are when an unbroken time series is plotted, the origin may be taken at the median value. When the number of observations is odd, this will be the middle item, of course. The value of $\Sigma(x)$ will then be zero, and the normal equations become

$$\Sigma(y) = na$$
$$\Sigma(xy) = b\Sigma(x^2).$$

Thus if a time series extends, by years, from 1900 to 1920, the origin may be taken at 1910, the value of x corresponding to 1909 being -1, to 1911, $+1$, and so on. The solution for values of a and b is rendered much easier when the data can be disposed in this way. When there is an even number of years the same process is possible, time (the x-variable) being measured in units of one half year.

Again, when the values of x are consecutive positive numbers starting at zero, the values of $\Sigma(x)$ and of $\Sigma(x^2)$ may be easily determined. The sum of the first n natural numbers is equal to $\frac{n(n+1)}{2}$. Thus the sum of the numbers from 1 to 10 is $\frac{10(10+1)}{2}$, or 55. This term may replace $\Sigma(x)$ in the normal equations. Similarly, the sum of the squares of the first n natural numbers is equal to $\frac{2n^3+3n^2+n}{6}$. Thus the sum of the squares of the numbers from 1 to 5 is equal to $\frac{250+75+5}{6} = 55$. This expression may replace $\Sigma(x^2)$ in the normal equations, and we have

$$\Sigma(y) = na + b\left(\frac{n(n+1)}{2}\right).$$

$$\Sigma(xy) = a\left(\frac{n(n+1)}{2}\right) + b\left(\frac{2n^3+3n^2+n}{6}\right).$$

It is sometimes easier to work from equations in this form than in the form first given. The data for time series may be handled in this way, the years being numbered consecutively, beginning with 1.

Fitting a Curve of the Potential Series

The discussion above has been confined to the case of linear trend. Such a type of curve frequently gives an excellent fit, but in many cases it fails accurately to fit the data. This difficulty is sometimes overcome in practice by breaking a series up into segments, and fitting a separate line to the data for each of these periods. Where there is an actual break in the series, the period as a whole lacking homogeneity, this practice may be justified, but when the period is essentially homogeneous the whole concept of secular trend is violated by this process of subdividing and fitting separate lines. In many cases where a straight line will not fit, a curve of the potential series may represent the trend accurately. The general process of fitting such a curve may be briefly described.

The generalized form of the equation of the type desired is $y = a + bx + cx^2 + dx^3 + \ldots$ An equation of this form does not, of course, represent a curve of the parabolic type, but in ordinary usage that term is applied to the potential series. If carried to the second power of x it is called a second degree parabola; if to the third power, a third degree parabola, etc. For ordinary purposes such a curve should not be carried beyond the second or third power of x. If carried to the second power there are three unknowns, and three normal equations must be solved simultaneously in securing the required values.

The procedure is similar to that outlined for the linear case. Each observation equation is multiplied by the coefficient of the first unknown in that equation, and the resulting equations are totaled to give the first normal

equation. The process is repeated for the two other unknowns, and the three normal equations thus obtained are solved for a, b and c. The results are the most probable values of these three constants. The following are the general forms which the three normal equations take:

$$\Sigma(y) = na + b\Sigma(x) + c\Sigma(x^2).$$
$$\Sigma(xy) = a\Sigma(x) + b\Sigma(x^2) + c\Sigma(x^3).$$
$$\Sigma(x^2y) = a\Sigma(x^2) + b\Sigma(x^3) + c\Sigma(x^4).$$

As an example of the process, the calculations involved in fitting a second degree parabola to the points 1, 2; 2, 6; 3, 7; 4, 8; 5, 10; 6, 11; 7, 11; 8, 10; 9, 9 may be outlined. It is of the greatest practical importance in curve fitting, as in all extensive calculations, that the work be laid out and carried on in a definite and systematic fashion, with each step definitely related to the preceding and succeeding operations. Checks should be introduced wherever possible, as mathematical errors creep into even the most careful work. A tabular arrangement is generally advisable, each operation being revealed and each set of results clearly presented.

The data in the present case may be arranged in the following form:

TABLE 68

Computation of Values Required in Fitting a Second Degree Parabola

x	y	xy	x^2	x^2y	
1	2	2	1	2	$n = 9$
2	6	12	4	24	$\Sigma(x) = 45$
3	7	21	9	63	$\Sigma(x^2) = 285$
4	8	32	16	128	$\Sigma(x^3) = 2{,}025$
5	10	50	25	250	$\Sigma(x^4) = 15{,}333$
6	11	66	36	396	$\Sigma(y) = 74$
7	11	77	49	539	$\Sigma(xy) = 421$
8	10	80	64	640	$\Sigma(x^2y) = 2{,}771$
9	9	81	81	729	
45	74	421	285	2,771	

When the x's are consecutive integers beginning with 1, as in the present case, the values of $\Sigma(x)$, $\Sigma(x^2)$, $\Sigma(x^3)$ and $\Sigma(x^4)$ may be secured from prepared tables.[1]

Substituting these values in the equations given above, the following normal equations are secured:

$$74 = 9a + 45b + 285c.$$
$$421 = 45a + 285b + 2{,}025c.$$
$$2{,}771 = 285a + 2{,}025b + 15{,}333c.$$

When these equations are solved simultaneously the following values are secured for the three constants:

$$a = -.929.$$
$$b = +3.523.$$
$$c = -.267.$$

The equation of the desired curve is

$$y = -.929 + 3.523x - .267x^2.$$

This curve and the nine given points are plotted in Fig. 56.

If the values of x are consecutive, as in the present example, the work of computation is lightened if the mid-value is taken as origin. In this case $\Sigma(x)$ and $\Sigma(x^3)$ are equal to zero, and the normal equations become

$$\Sigma y = na + c\Sigma(x^2).$$
$$\Sigma(xy) = b\Sigma(x^2).$$
$$\Sigma(x^2y) = a\Sigma(x^2) + c\Sigma(x^4).$$

When a third degree parabola of the form $y = a + bx + cx^2 + dx^3$ is to be fitted to data, four constants must be determined, and four normal equations are necessary. These are of the following form:

$$\Sigma(y) = na + b\Sigma(x) + c\Sigma(x^2) + d\Sigma(x^3).$$
$$\Sigma(xy) = a\Sigma(x) + b\Sigma(x^2) + c\Sigma(x^3) + d\Sigma(x^4).$$
$$\Sigma(x^2y) = a\Sigma(x^2) + b\Sigma(x^3) + c\Sigma(x^4) + d\Sigma(x^5).$$
$$\Sigma(x^3y) = a\Sigma(x^3) + b\Sigma(x^4) + c\Sigma(x^5) + d\Sigma(x^6).$$

[1] *Cf.* Table XXVIII, Pearson, *Tables for Statisticians and Biometricians.* Cambridge University Press, 1914.

The solution for four or more constants involves a considerable amount of arithmetical calculation, and there is some question as to the advisability of representing secular trend by equations of this type. With a sufficient number of constants a curve may be fitted which will follow every

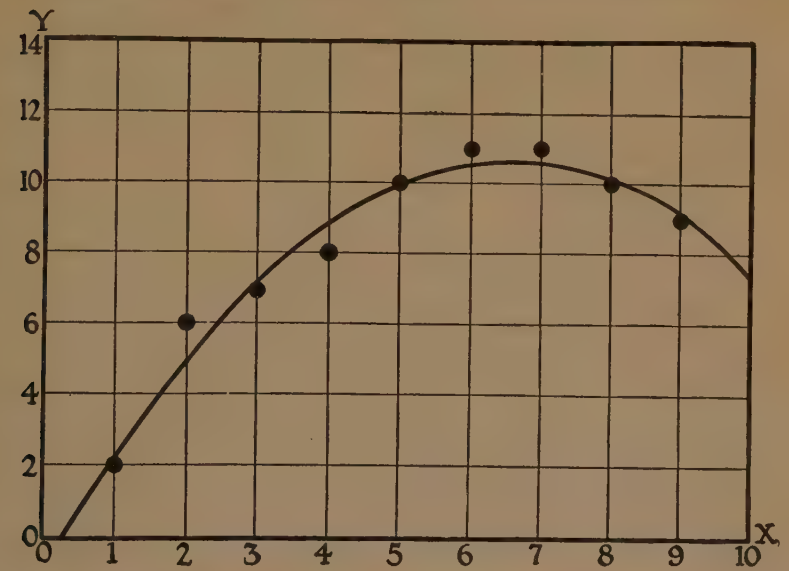


FIG. 56. — Illustrating the Fitting of a Second Degree Parabola to Nine Points

variation in the data, but such a curve could hardly be taken to represent the long time trend.[1] Minor departures from a simple and uniform trend, linear or otherwise, are to be expected with economic data, but, if a real trend exists,

[1] Regarding the employment of potential series of the type indicated for representing empirical curves, Steinmetz states that their use is justified:

1. If the successive coefficients *a*, *b*, *c* . . . decrease in value so rapidly that within the range of observation the higher terms become rapidly smaller and appear as mere secondary terms.
2. If the successive coefficients follow a definite law, indicating a convergent series which represents some other function, as an exponential, trigonometric, etc.
3. If all the coefficients are very small, with the exception of a few of them, and only the latter ones thus need to be considered. *Cf.* Steinmetz, *Engineering Mathematics*, 214–215.

extreme departures from a fairly simple form are rare. If such departures are due to pronounced changes in conditions no single line of trend is likely to be satisfactory, and it is advisable to break the period up into parts, with a separate line of trend for each part. "Empirical curves," says Steinmetz, "can be represented by a single equation only when the physical conditions remain constant within the range of the observations." Though this statement relates to the fitting of curves to data from the physical sciences, it applies equally well to economic data.

A Typical Problem: Determining the Secular Trend of Business Failures

The procedure of fitting certain types of curves to simple data has been illustrated in the preceding sections. Before proceeding to a discussion of slightly different forms, it may be helpful to insert a concrete example at this point. A straight line, a second degree parabola and a third degree parabola are to be fitted to the figures for business failures in the United States, from 1897 to 1921. The three curves are to be fitted in order that the results may be compared and significant differences noted.

To facilitate the calculations, the mid-year of the period, 1909, is taken as the origin. Certain of the values needed in the normal equations are computed in the following table. The values of x represent the time factor, while the values of y are the corresponding numbers of business failures:

TABLE 69

Business Failures in the United States, 1897–1921

Computation of Values Required in Fitting Lines of Trend

(1)	(2)	(3)	(4)	(5)	(6)
Year	x	y (*No. of failures*)	xy	x^2y	x^3y
1897	− 12	13,083	− 156,996	1,883,952	− 22,607,424
1898	− 11	11,615	− 127,765	1,405,415	− 15,459,565
1899	− 10	9,642	− 96,420	964,200	− 9,642,000
1900	− 9	9,912	− 89,208	802,872	− 7,225,848
1901	− 8	10,648	− 85,184	681,472	− 5,451,776
1902	− 7	9,973	− 69,811	488,677	− 3,420,739
1903	− 6	9,775	− 58,650	351,900	− 2,111,400
1904	− 5	10,417	− 52,085	260,425	− 1,302,125
1905	− 4	9,967	− 39,868	159,472	− 637,888
1906	− 3	9,385	− 28,155	84,465	− 253,395
1907	− 2	10,274	− 20,548	41,096	− 82,192
1908	− 1	14,066	− 14,066	14,066	− 14,066
1909	0	11,872	.....	.....	.....
1910	1	11,588	11,588	11,588	11,588
1911	2	12,679	25,358	50,716	101,432
1912	3	13,832	41,496	124,488	373,464
1913	4	14,553	58,212	232,848	931,392
1914	5	16,780	83,900	419,500	2,097,500
1915	6	19,035	114,210	685,260	4,111,560
1916	7	16,498	115,486	808,402	5,658,814
1917	8	13,073	104,584	836,672	6,693,376
1918	9	9,331	83,979	755,811	6,802,299
1919	10	5,515	55,150	551,500	5,515,000
1920	11	8,463	93,093	1,024,023	11,264,253
1921	12	19,982	239,784	2,877,408	34,528,896
Totals	0	301,958	+ 188,084	15,516,228	+ 9,881,156

$n = 25$

$\Sigma(x) = 0$

$\Sigma(y) = 301{,}958$

$\Sigma(xy) = 188{,}084$

$\Sigma(x^2) = 1{,}300$

$\Sigma(x^2y) = 15{,}516{,}228$

$\Sigma(x^3) = 0$

$\Sigma(x^3y) = 9{,}881{,}156$

$\Sigma(x^4) = 121{,}420$

$\Sigma(x^5) = 0$

$\Sigma(x^6) = 13{,}471{,}900$

Fitting a Straight Line to Business Failures

Since the origin is at the mid-point of the x's, the equations to be solved in securing the values of the required constants are of the form

$$\Sigma(y) = na$$
$$\Sigma(xy) = b\Sigma(x^2).$$

Inserting the given values in the formulas, we have

$$301{,}958 = 25a$$
$$188{,}084 = 1{,}300b$$

from which

$$a = 12{,}078$$
$$b = 144.7.$$

The equation of the line of best fit is, therefore,

$$y = 12{,}078 + 144.7x.$$

Fitting a Second Degree Parabola

For a second degree parabola the normal equations to be solved are of the form

$$\Sigma(y) = na + c\Sigma(x^2)$$
$$\Sigma(xy) = b\Sigma(x^2)$$
$$\Sigma(x^2y) = a\Sigma(x^2) + c\Sigma(x^4).$$

When the appropriate values are substituted, the following equations are secured:

$$301{,}958 = 25a + 1{,}300c$$
$$188{,}084 = 1{,}300b$$
$$15{,}516{,}228 = 1{,}300a + 121{,}420c.$$

Solving for the constants,

$$a = +12{,}258$$
$$b = +144.7$$
$$c = -3.45.$$

The required equation is:

$$y = 12{,}258 + 144.7x - 3.45x^2.$$

Fitting a Third Degree Parabola

The normal equations to be solved in securing the values of the constants for a third degree parabola, when the origin is at the mid-point of the x's, are of the form

$$\begin{aligned}
\Sigma(y) &= na + c\Sigma(x^2) \\
\Sigma(xy) &= b\Sigma(x^2) + d\Sigma(x^4) \\
\Sigma(x^2y) &= a\Sigma(x^2) + c\Sigma(x^4) \\
\Sigma(x^3y) &= b\Sigma(x^4) + d\Sigma(x^6).
\end{aligned}$$

The equations to be solved simultaneously are:

$$\begin{aligned}
301{,}958 &= 25a + 1{,}300c \\
188{,}084 &= 1{,}300b + 121{,}420d \\
15{,}516{,}228 &= 1{,}300a + 121{,}420c \\
9{,}881{,}156 &= 121{,}420b + 13{,}471{,}900d.
\end{aligned}$$

The values derived from these equations are:

$$\begin{aligned}
a &= +12{,}258 \\
b &= +482 \\
c &= -3.45 \\
d &= -3.61.
\end{aligned}$$

The required equation is:

$$y = 12{,}258 + 482x - 3.45x^2 - 3.61x^3.$$

The original data and the lines of trend given by these three equations are plotted in Fig. 57. The actual values of y (business failures), the computed or *normal* values in each of the three cases and the percentage deviations of actual from normal in each case are shown in the following table:

TABLE 70

Business Failures in the United States, 1897–1921

Actual Values, Normal Values as derived from three lines of trend, and Percentage Deviations of Actual from Normal

Year	Actual Values	Normal Values Straight Line	Normal Values Second Degree Parabola	Normal Values Third Degree Parabola	Deviations		
					From Straight Line	From Second Degree Parabola	From Third Degree Parabola
1897	13,083	10,342	10,025	12,215	+ 26.5%	+ 30.5%	+ 7.0%
1898	11,615	10,486	10,249	11,344	+ 10.8	+ 13.2	+ 2.3
1899	9,642	10,631	10,466	10,703	− 9.3	− 8.0	− 10.0
1900	9,912	10,776	10,676	10,272	− 8.0	− 7.0	− 3.8
1901	10,648	10,920	10,880	10,030	− 2.5	− 2.0	+ 6.5
1902	9,973	11,065	11,076	9,953	− 9.9	− 10.0	+ 0.2
1903	9,775	11,210	11,266	10,022	− 12.8	− 13.2	− 2.5
1904	10,417	11,354	11,448	10,213	− 8.3	− 9.0	+ 2.1
1905	9,967	11,499	11,624	10,506	− 13.3	− 14.3	− 5.1
1906	9,385	11,644	11,793	10,878	− 19.4	− 20.4	− 13.7
1907	10,274	11,789	11,955	11,313	− 12.9	− 14.1	− 9.2
1908	14,066	11,933	12,110	11,776	+ 17.8	+ 16.0	+ 19.4
1909	11,872	12,078	12,258	12,258	− 1.7	− 3.2	− 3.2
1910	11,588	12,223	12,400	12,733	− 5.2	− 6.6	− 9.1
1911	12,679	12,367	12,533	13,179	+ 2.5	+ 1.0	− 3.8
1912	13,832	12,512	12,660	13,575	+ 10.5	+ 9.2	+ 1.9
1913	14,553	12,657	12,781	13,900	+ 15.0	+ 13.9	+ 4.6
1914	16,780	12,801	12,895	14,131	+ 31.0	+ 30.0	+ 18.6
1915	19,035	12,946	13,002	14,246	+ 47.0	+ 46.2	+ 33.5
1916	16,498	13,091	13,102	14,225	+ 26.0	+ 25.9	+ 15.9
1917	13,073	13,236	13,194	14,045	− 1.2	− 0.9	− 7.0
1918	9,331	13,380	13,281	13,685	− 30.2	− 29.8	− 31.8
1919	5,515	13,525	13,360	13,123	− 59.1	− 58.7	− 57.8
1920	8,463	13,670	13,432	12,338	− 38.0	− 37.0	− 31.4
1921	19,982	13,814	13,498	11,307	+ 44.6	+ 48.0	+ 76.8

COMPARISON OF LINES OF TREND

It is clear from the preceding discussion, and from Fig. 57, that widely different results may be secured when the secular trend in a given case is represented by different types of curves. In the present instance the straight line and the second degree parabola follow much the same course, but the third degree parabola marks out quite a different trend. The example was selected for the purpose of emphasizing these differences. Since a "normal" for each year is measured by the ordinate of the line of trend for that year, the use of different lines of trend may give quite

different standards of normality. Moreover, the cycles registered by the deviations of the actual data from the line of trend may vary with the type of curve selected. That this is true in the present case is shown by the percentage deviation figures in Table 70.

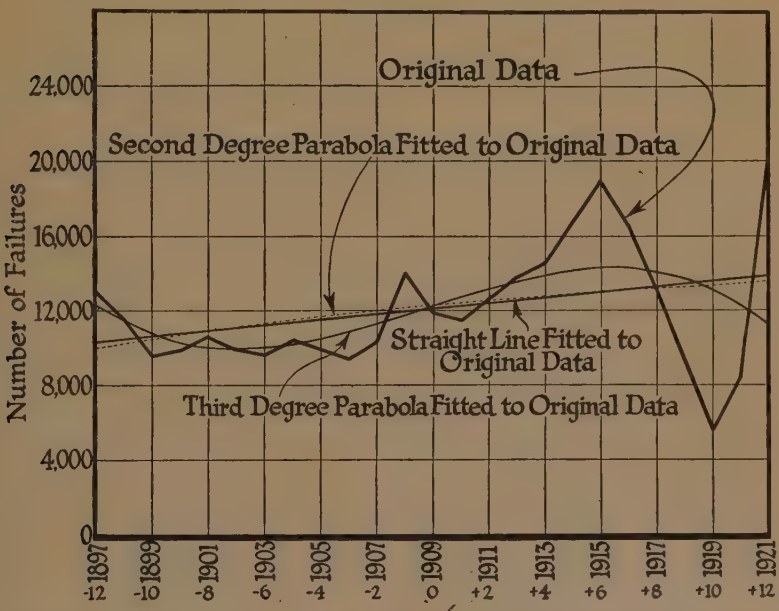


FIG. 57. — Business Failures in the United States, 1897–1921, with Three Lines of Trend

The deviations from the straight line and the second degree parabola show a close correspondence, but the deviations from the third degree parabola are markedly different. In four years the signs of the deviations from the third degree parabola are different from those of the deviations from the other curves, and in many of the other years the magnitudes of the deviations vary widely. An extension of this third degree curve beyond the limits of the data would give quite illogical results, however. While projection in any case is risky, the straight line and the

second degree parabola constitute measures of trend which are better for this purpose than the third degree parabola.

More will be said in a later section as to the factors which determine the choice of a curve. With respect to the selection of a curve in the present instance several points should be noted. If the actual data fall consistently above or below a line of trend for a considerable period, it is probable that the fit is not good. For business failures both the straight line and the second degree parabola fail to follow the actual data closely, there being a single nine-year period from 1899 to 1907 during which the data are continually below normal, as represented by these two lines of trend. The third degree parabola follows the data much more closely, and, so far as one may judge by inspection, gives a better fit within the limits of the data.

The Use of Logarithms in Curve Fitting

The family of curves described above represents a simple and very useful type. Perhaps of even greater general utility, in the analysis of time series, are curves of a semi-logarithmic type. The advantages of plotting many series of data on semi-logarithmic or "ratio" paper were explained in an earlier section. A fundamental virtue of this type of plotting is that it presents a true picture of *relative* variations, of *ratios* between magnitudes. Relations of this type are ordinarily of primary interest in the analysis of economic data, and therefore the determination of trends should proceed on the same basis.

In doing so, we can make use of a group of curves of the same general form as those already described, the one difference being that $\log y$ takes the place of y throughout. That is, the straight line form is $\log y = a + bx$, while the general form for the potential series is $\log y = a + bx + cx^2 + dx^3 + \ldots$ The curves secured may be constructed on arithmetic paper, plotting the natural x's and the logarithms

of the y's, or natural values of both x's and y's may be plotted on semi-logarithmic paper, the logarithmic scale extending along the y-axis. The latter is the simpler method.

To illustrate the procedure, the steps involved in fitting a curve of the type $\log y = a + bx + cx^2$ will be shown. The trend of petroleum production in the United States from 1908 to 1922 is to be determined. The values needed in the normal equations are derived from the following table:

TABLE 71

Petroleum Production in the United States, 1908–1922

Computation of Values required in fitting Line of Trend

Year	x	y *Production (in millions of bbls.)*	*log y*	$x \cdot \log y$	$x^2 \cdot \log y$
1908	− 7	178.5	2.25164	− 15.76148	110.33036
1909	− 6	183.2	2.26269	− 13.57614	81.45684
1910	− 5	209.6	2.32139	− 11.60695	58.03475
1911	− 4	220.4	2.34321	− 9.37284	37.49136
1912	− 3	222.9	2.34811	− 7.04433	21.13299
1913	− 2	248.4	2.39515	− 4.79030	9.58060
1914	− 1	265.8	2.42456	− 2.42456	2.42456
1915	0	281.1	2.44886	.....	.....
1916	1	300.8	2.47828	2.47828	2.47828
1917	2	335.3	2.52543	5.05086	10.10172
1918	3	355.9	2.55133	7.65399	22.96197
1919	4	378.4	2.57795	10.31180	41.24720
1920	5	442.9	2.64631	13.23155	66.15775
1921	6	472.2	2.67413	16.04478	96.26868
1922	7	557.5	2.74624	19.22368	134.56576
			36.99528	9.41834	694.23282

$n = 15$ $\qquad \Sigma(x) = 0$

$\Sigma(\log y) = 36.99528 \qquad \Sigma(x^2) = 280$

$\Sigma(x \cdot \log y) = 9.41834 \qquad \Sigma(x^3) = 0$

$\Sigma(x^2 \cdot \log y) = 694.23282 \qquad \Sigma(x^4) = 9352$

The three normal equations to be solved are of the form

$$\Sigma(\log y) = na + b\Sigma x + c\Sigma x^2$$

$$\Sigma(x \cdot \log y) = a\Sigma x + b\Sigma x^2 + c\Sigma x^3$$

$$\Sigma(x^2 \cdot \log y) = a\Sigma x^2 + b\Sigma x^3 + c\Sigma x^4.$$

Substituting the given values we have

$$36.99528 = 15a + 280c$$
$$9.41834 = 280b$$
$$694.23282 = 280a + 9352c.$$

Solving for the constants,

$$a = +2.450508$$
$$b = + .033637$$
$$c = + .0008488.$$

The equation to the desired curve is, therefore,

$$\log y = 2.450508 + .033637x + .0008488x^2,$$

with origin at 1915.

The substitution in this equation of the value of x representing any given year will enable the logarithm of the trend or normal value to be calculated. The trend value in natural numbers may then be determined. In the table which follows the normal value for each of the years covered is given, together with the percentage relation of actual to normal.

TABLE 72

Trend of Petroleum Production in the United States with Comparison of Actual and Trend Values

(Line of Trend fitted to Logarithms of Production Figures.)

Year	x	y (*actual*) *Production (in millions of bbls.)*	*Log of Trend*	y (*computed*) *Trend value (in millions of bbls.)*	*Percentage relation of actual to trend*
1908	−7	178.5	2.256639	180.6	98.8
1909	−6	183.2	2.279242	190.2	96.3
1910	−5	209.6	2.303543	201.2	104.2
1911	−4	220.4	2.329540	213.6	103.2
1912	−3	222.9	2.357236	227.6	97.9
1913	−2	248.4	2.386629	243.6	102.0
1914	−1	265.8	2.417720	261.6	101.6
1915	0	281.1	2.450508	282.2	99.6
1916	1	300.8	2.484994	305.5	98.5
1917	2	335.3	2.521177	332.0	101.0
1918	3	355.9	2.559058	362.3	98.2
1919	4	378.4	2.598636	396.9	95.3
1920	5	442.9	2.639913	436.4	101.5
1921	6	472.2	2.682886	481.8	98.0
1922	7	557.5	2.727557	534.0	104.4

The points representing the actual production, together with the line of trend, are plotted in Fig. 58. The graph of the derived equation gives a very good representation of the trend in the present instance.

The advantages of this type of curve are evident when the above results are compared with those secured when a

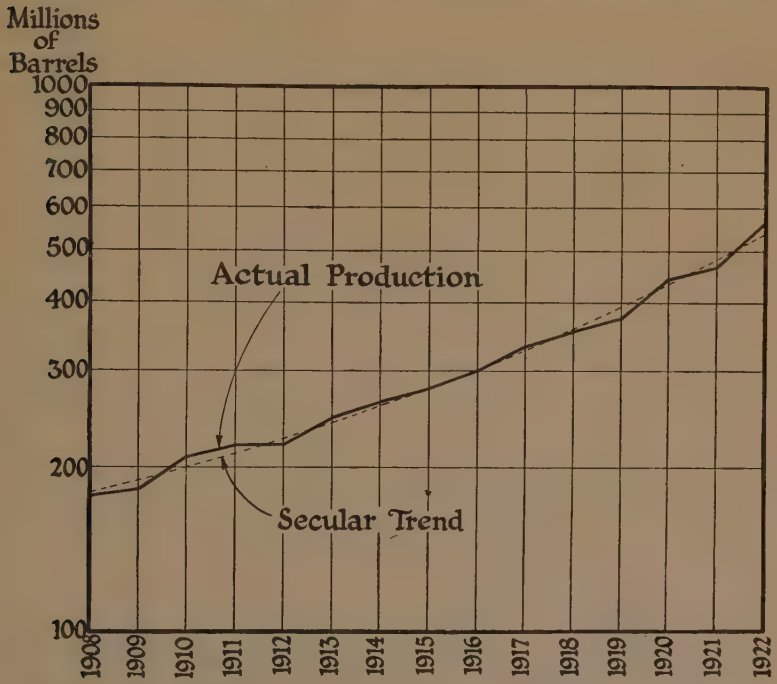


FIG. 58. — Production of Crude Petroleum in the United States, 1908–1922, with Line of Trend Fitted to Logarithms

similar curve (a second degree parabola) is fitted to the natural numbers. The values of the constants are secured by the method of least squares, and the following equation is derived:

$$y = 279.387 + 24.262x + 1.650x^2$$

with origin at 1915. The trend values secured from this equation, together with the actual production figures and the percentage relations between them, are given below.

TABLE 73

Trend of Petroleum Production in the United States with Comparison of Actual and Trend Values

(Line of Trend fitted to Natural Numbers.)

Year	*x*	*y (actual) Production (in millions of bbls.)*	*y (computed) Trend value (in millions of bbls.)*	*Percentage relation of actual to trend*
1908	− 7	178.5	190.403	93.7
1909	− 6	183.2	193.215	94.8
1910	− 5	209.6	199.327	105.2
1911	− 4	220.4	208.739	105.6
1912	− 3	222.9	221.451	100.7
1913	− 2	248.4	237.463	104.6
1914	− 1	265.8	256.775	103.5
1915	0	281.1	279.387	100.6
1916	1	300.8	305.299	98.5
1917	2	335.3	334.511	100.2
1918	3	355.9	367.023	97.0
1919	4	378.4	402.835	93.9
1920	5	442.9	441.947	100.2
1921	6	472.2	484.359	97.5
1922	7	557.5	530.071	105.2

This curve, together with the actual production figures, is plotted in Fig. 59.

Each of the equations used to represent the trend of petroleum production in this example contains three constants, so comparison of the results is legitimate. Inspection of the two graphs would lead to the conclusion that the curve fitted to the logarithms of the data gives a better representation of the trend than the curve fitted to the natural numbers, but an accurate measure is preferable to such a general impression. Such a measure is afforded by the root-mean-square deviation. This may be computed exactly as in dealing with a frequency distribution, except that deviations are measured from the line of trend instead of from the arithmetic mean.

In the case of the curve fitted to the natural numbers, the root-mean-square (standard) deviation is found to be 12.46, while for the curve fitted to logarithms of the pro-

duction figures this measure has a value of 9.44. (The unit in each case is one million barrels of petroleum.) It is clear that the dispersion about the latter curve is distinctly less than about the curve based on natural numbers;

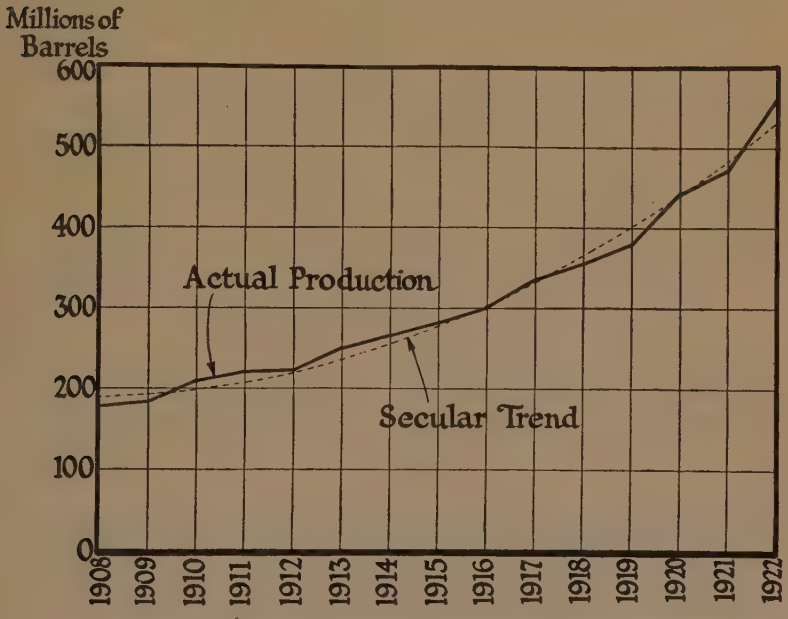


FIG. 59. — Production of Crude Petroleum in the United States, 1908–1922, with Line of Trend Fitted to Natural Numbers

the logarithmic curve is by far the better measure of trend in the present instance.[1]

[1] The standard deviations given above were computed from the actual deviations, in millions of barrels of petroleum. The superiority of the logarithmic curve is even more pronounced, as would be expected, if these measures are computed from percentage deviations. The standard deviation about the curve fitted to the natural numbers is 4.01 on this basis, while the standard deviation about the logarithmic curve is but 2.69.

It should be noted that the testing of lines of trend by comparing standard deviations is only valid when the equations to the curves being compared contain the same number of constants. A curve may be made to pass through every point, if a number of constants be introduced equal to the number of points plotted. The standard deviation in this case would be zero, but such a curve would not be in any sense a measure of trend.

The two families of curves described in the preceding sections meet most of the needs of the economic statistician. The trend in most time series may be described by curves of the potential series, fitted either to natural numbers or to the logarithms of the data (that is, to the logarithms of the y values; time, the x-variable, is treated in terms of natural numbers in fitting both the above types of curves). These classes constitute flexible and widely applicable curve forms. Attention may be called to several other curve types which have been applied less extensively to time series, but with favorable results in particular cases.

Curves of the ordinary parabolic type ($y = ax^b$) are not generally applicable to economic data in the form of time series, as their use involves the treatment of the time variable as a geometric series. Such a curve, it will be recalled, becomes a straight line on double logarithmic paper. Yet if a curve of this form serves accurately to describe the trend of a given series, its use is justified, empirically.

Such curves may be fitted most readily by employing logarithms and using an equation of the linear type. The equation

$$y = ax^b$$

becomes, in logarithmic form,

$$\log y = \log a + b \log x.$$

The two normal equations needed in fitting such a curve are of the form

$$\Sigma(\log y) = n \log a + b\Sigma(\log x)$$
$$\Sigma(\log x \cdot \log y) = \log a\Sigma(\log x) + b\Sigma(\log x)^2.$$

By substituting the values computed from the data, these equations may be solved for $\log a$ and b, just as in fitting an ordinary straight line.[1]

[1] A very useful table of the sums of the logarithms of the natural numbers from 1 to 100 is included as an appendix to *Medical Biometry and Statistics*, by Raymond Pearl. (Philadelphia, Saunders, 1923.)

The simple exponential curve ($y = ab$) may be widely used in the analysis of time series. In logarithmic form this becomes

$$\log y = \log a + (\log b)x.$$

The two normal equations to be solved in fitting such a curve are

$$\Sigma(\log y) = n \log a + \log b \Sigma(x)$$
$$\Sigma(x \cdot \log y) = \log a \Sigma(x) + \log b \Sigma(x^2).$$

This is the straight line form of the family of semi-logarithmic curves described in an earlier section. By determining the antilogarithms of the two constants secured from the solution of the normal equations, the derived equation may be put in the exponential form.

Some use has been made, in the interpretation of economic statistics, of the Gompertz curve, the equation to which was originally developed in the actuarial field. The equation is

$$y = ab^{c^x}.$$

Its use in the analysis of economic statistics has been based upon the argument that there is a general law of growth characteristic of population increase, and that this same type of growth is found in industries whose products are a direct function of the growth of population.[1]

A somewhat similar curve of growth described by the equation

$$y = d + \frac{b}{e^{-ax} + c}$$

has been employed by Raymond Pearl and Lowell J. Reed in forecasting population growth.[2] This curve has been found to describe the trend of certain economic data.

[1] The characteristics of this curve are described in "Law of Growth in Forecasting Demand," by Raymond B. Prescott, *Journal of the American Statistical Association*, Dec. 1922, 471–79.

[2] "Predicted Growth of Population of New York and Its Environs." R. Pearl and L. J. Reed. Committee of Plan of New York and Its Environs, 1923.

Short Cut Methods of Fitting Curves: the Method of Averages

The fitting of curves by the method of least squares is appropriate when a high degree of accuracy is required. Other methods of fitting curves may sometimes be employed when only approximate values are required.

In fitting a straight line by the method of averages the series is divided into two equal or nearly equal parts. From the observations falling in the two groups two equations of the following form are secured:

$$\Sigma(y) = na + b\Sigma(x).$$

These equations are solved simultaneously, and the values of a and b obtained. This is equivalent to finding, for each half of the series, the point having coördinates equal to the arithmetic averages of the x's and y's, and connecting the two points thus secured. This line will not be identical, ordinarily, with the line of least squares.

The method may be illustrated by using the data of business failures from Table 69. Since the series covers 25 years, 13 observations may be included in the first group and 12 in the second. The two equations are

$$140{,}629 = 13a - 78b$$
$$161{,}329 = 12a + 78b.$$

Solving,

$$a = 12{,}078$$
$$b = 210$$

and the equation to the line of trend is

$$y = 12{,}078 + 210x.$$

By the method of least squares the equation

$$y = 12{,}078 + 144.7x$$

was obtained. The latter, of course, is the equation to the best fitting straight line. The short method in this case gives an equation varying considerably from the correct

result. This may be expected when the original data depart as widely from the line of trend as in the present case.

The Method of Selected Points

Having determined the form of curve to be fitted to certain data, approximate values of the constants may be secured by selecting a number of points equal to the number of constants required. The coördinates of these points are substituted in the corresponding number of equations, and these are solved for the constants. Thus, if a third degree parabola is to be fitted to the data of business failures, four points must be selected. Drawing an experimental curve by inspection will assist in the selection of appropriate points. The following points may be taken as a basis for calculation (*cf.* Table 69 and Fig. 57 relating to business failures).

x	y
− 10	10,000
− 5	10,000
+ 5	14,000
+ 10	13,000

Substituting these values for x and y in equations of the form desired, the following are secured:

$$10{,}000 = a - 10b + 100c - 1{,}000d$$
$$10{,}000 = a - 5b + 25c - 125d$$
$$14{,}000 = a + 5b + 25c + 125d$$
$$13{,}000 = a + 10b + 100c + 1{,}000d.$$

Solving these equations simultaneously, the following values are obtained for the four constants:

$$a = + 12{,}166$$
$$b = + 483$$
$$c = - 6.7$$
$$d = - 3.3.$$

The required equation is, therefore,

$$y = 12{,}166 + 483x - 6.7x^2 - 3.3x^3.$$

The difference between this equation and the one secured by the method of least squares will be noted.

There is a considerable margin of error in the employment of these short methods with ordinary economic data, because of the fact that most economic series do not adhere closely to a given line of trend. That is, economic data correspond to physical measurements characterized by large errors of observation. When physical data follow a definite law, and the errors of observation are small, the use of the method of averages or the method of selected points will involve no material error. But with economic data, which frequently depart widely from any given line of trend, the use of short cut methods is in general inadvisable.

The Selection of a Curve to Represent Trend

Various types of curves which may be fitted to represent the trend of economic data over a period of time have been described. But which of these many types is to be selected in a given case? Which will give the best standard of normality for each of the years covered? Several references to this problem have been made in the preceding sections, but no general principles have been laid down. And, in fact, no general principles can be evoked to answer this fundamental question. There is no absolute test of goodness of fit in such cases. It is largely a matter of personal judgment as to the type of curve which best represents the trend in a given instance, and experience must play a dominant part in such judgments. But there are certain general considerations which are of assistance in selecting the appropriate type of curve, and which, in some cases, enable a single type to be selected with confidence that it is the best.

1. The first step in the selection of a curve type is the plotting of the data. When this has been done, it is fre-

quently possible by inspection to determine the appropriate form. The data may be plotted in four different combinations, of which the first two are of chief importance in dealing with economic material.

a. Natural x, natural y. (That is, plot the given figures on ordinary arithmetic paper.)
b. Natural x, log y. (Plot the x's on the natural scale, and plot the y's on the logarithmic scale; i.e., use semi-logarithmic paper.)
c. Natural y, log x. (Plot on semi-logarithmic paper, with the x-scale logarithmic.)
d. Log y, log x. (Plot on double logarithmic paper.)

If in any of these cases a straight line trend is secured, a type of equation which plots as a straight line under the given conditions (*cf.* Chapter II) would be selected. If a linear equation cannot be secured some other simple type may be suggested by the plotted data. In studying such graphs for the purpose of selecting a curve to represent trend, one should be familiar with the curves representing all the simpler equations.[1]

2. The appropriate curve may be determined by a study of the relations between the two variables, x and y. In the simplest cases the following relations hold:[2]

If, when the values of x are arranged in an arithmetic series, the corresponding values of y form a geometric series, the relation is of the exponential type, described by the equation

$$y = ab^x$$

[1] A variety of curves will be found plotted in the following:
Empirical Formulas, T. R. Running, N. Y., Wiley, 1917.
Engineering Mathematics, C. P. Steinmetz, N. Y., McGraw-Hill, 1917.
Graphical and Mechanical Computation, Joseph Lipka, N. Y., Wiley, 1918.

[2] It will be recalled in connection with this discussion that an arithmetic series changes by a constant absolute increment, while a geometric series changes by a constant percentage.

b. If, when the values of x are arranged in a geometric series, the corresponding values of y form a geometric series the relation is of the simple parabolic or hyperbolic type, described by the equation

$$y = ax^b$$

c. If, when the values of x are arranged in an arithmetic series, the first differences of the corresponding y's are constant, the relation is of the straight line type, described by the equation

$$y = a + bx$$

The differences between successive y values, when x's are arranged in an arithmetic series, are termed "first differences" or first order "differences" and are represented by the symbol Δy. The differences between successive first differences are called "second differences" and are represented by the symbol $\Delta^2 y$. Differences of higher order are similarly derived. The following table illustrates the formation of differences:

x	y	Δy	$\Delta^2 y$	$\Delta^3 y$
1	11			
		29		
2	40		32	
		61		12
3	101		44	
		105		12
4	206		56	
		161		12
5	367		68	
		229		12
6	596		80	
		309		12
7	905		92	
		401		12
8	1306		104	
		505		12
9	1811		116	
		621		
10	2432			

d. If, when the values of x are arranged in an arithmetic series, the nth differences of the corresponding y's are constant, the relation between the variables is described by an equation of the potential series carried to the nth power of x; that is, by an equation of the type

$$y = a + bx + cx^2 + dx^3 + \ldots qx^n.$$

Thus, in the example given above, in which the third differences are constant, the relation between x and y would be described by an equation of the form

$$y = a + bx + cx^2 + dx^3.$$

These and similar tests, which may be applied to a great many other types of data, are described in the books by Running and Lipka referred to at the end of this chapter. It should be emphasized that, when one is selecting a curve to use in the analysis of economic data, he will rarely, if ever, find these tests to be met perfectly. This would happen only when the curve chosen passed through all the plotted points. But data in a given case will generally approximate some one of the conditions described above, and the appropriate type of curve will be indicated.

3. If the study of the original data does not render a definite decision possible, several types of curves may be fitted to the data and the decision made by comparing the results, as was done in the two cases already cited, business failures and petroleum production. If the equations to the curves being compared contain the same number of constants, a comparison of the root-mean-square deviations about the curves furnishes a conclusive and valid test of the closeness of the fit within the limits of the data.

The root-mean-square deviation may be readily computed by making use of the following relationship

$$\Sigma(d^2) = \Sigma(y^2) - a\Sigma(y) - b\Sigma(xy) - c\Sigma(x^2y) - \ldots$$

where $\Sigma\ (d^2)$ is the sum of the squares of the deviations about the line of trend. (The derivation of this equation is explained in Appendix A, in which a generalized form is given.) If the equations do not contain an equal number of constants, a test of this sort is invalid and the comparison can only be made by inspection. Personal judgment as to the curve which represents the trend most accurately must be the basis of the decision in such cases.

It should be remembered that the closeness of fit within the limits of the data is not by itself a final criterion. An equation could be secured, having a number of constants equal to the number of points, which would give a curve passing through every point plotted, yet such a curve

would not necessarily represent the trend. The concept of a *trend* is of a regular, smooth underlying movement, from which there are deviations, but which marks the long time tendency of the series. In general, therefore, the curve should be of simple form, if it is to be consistent with the concept of secular trend. This does not mean, however, that a complex trend can be represented by a simple curve which fails to conform to the plotted data.

4. An important question to be answered before the form of curve can be selected relates to the limits within which the line of trend is to be used. If it is to be used only within the limits of the plotted data (i.e., for *interpolation*) one set of considerations governs the choice of a curve. If it is to be projected beyond the limits of the data, used as a basis for the determination of *normal* during a subsequent period, other considerations enter. In the former case a reasonable fit to the data is the sole requirement; in the latter case it is necessary, in addition, that the trend of the projection be logical, and consistent with the past record. This requirement was emphasized in discussing the trend of business failures.

The fact should be clearly recognized that projection, or *extrapolation*, represents a guess, justified only on the assumption that a proper line of trend has been fitted and that the same conditions which affected the series in the past will prevail in the future. A change in conditions, the introduction of new elements, renders the projection invalid. When dealing with economic statistics, moreover, it is ordinarily impossible to tell, except in retrospect, when a change has taken place. Conclusions drawn from the projection of a line of trend are always subject to error, therefore. In practical statistical work such projections are made, and are justified on the ground that the most probable course in the future is that which prevailed in the past. Projections into the distant future are, of course, subject to wider margins of error than short-time projections.

Lines of trend should be revised from time to time, therefore, as new data become available.

When a projection is to be made a simple curve with few constants is to be preferred to a more complicated one. A third or fourth degree parabola may give an excellent fit to the data in a given case, but the projection of such curves is inadvisable. It is well to remember, as Perrin has pointed out, that a curve suitable for interpolation may not be at all adapted to extrapolation.

It seems to be true, in general, that simple curves fitted to the logarithms of the y's give more reliable results when projected than curves fitted to the natural numbers. In an interesting discussion of this point, Karl G. Karsten [1] argues that phenomena characterized by a uniform *rate of change* are more likely to maintain their trend than phenomena marked by a uniform *amount of change*. It is the semi-logarithmic curves, of course, which best measure *rates* of change.

5. When the object of fitting a line of trend is the study of deviations connected with business cycles, another test of the accuracy of fit is afforded by a comparison of the deviation cycles with actual business conditions, as measured by a number of other series. In general, that line of trend which gives cycles consistent with the cycles in general business is to be preferred. The curves fitted to the data of business failures may be tested in this way. The broad cycles through which business moved during the period covered are reflected in the deviations from the straight line and the second degree parabola, but the minor cycles are not. These two lines of trend show business failures to have been below normal during the whole period from 1899 to 1907. When the trend is represented by the third degree parabola, failures slightly above normal are found in 1901 and in 1904. These represent phases of depression in minor cycles, though the whole period from 1899 to 1907 was one

[1] *Charts and Graphs*, 423–425.

of general prosperity. The curve of higher degree, it is apparent, follows the data more closely, registering minor deviations as cycles. Whether this is desirable in a given case depends upon the nature of the study being made.

6. It is frequently true that no one curve will fit a given series during the entire period it is desired to study. This may be due to changes in conditions which cause the trend to be altered. Thus the trend of wholesale prices was downward, in a direction well represented by a straight line, from the close of the Civil War to 1896. From 1896 to the beginning of the World War the trend was upward, and could be described by a second degree parabola. Similar changes occur in many economic series. By breaking the entire period up into sections, appropriate lines of trend may be fitted to the several periods thus marked off. This process may be carried to a quite illogical extreme, however. The concept of trend is of a gradual, long-term change, and the breaking up of a series in order to fit a number of trend lines is contrary to the whole conception. It may be justified upon occasion when a real change in conditions occurs, but in all cases the attempt should be made to represent the trend during the whole period by a single line.

Representation of Trend by a Related Statistical Series

It is sometimes impossible to represent the trend of a series by a mathematical curve. This is likely to be the case when data are available over but a short period, or when the trend of an old series has been quite clearly changed by the introduction of new factors. Such a problem is presented by commodity prices since the war. The pre-war trend has been definitely altered, and it is impossible to say with any degree of certainty what the future trend will be. Faced with this difficulty, Warren M. Persons has

developed a novel method of measuring cyclical fluctuations by using one series as the base of a related series.[1]

In the field of prices the two series employed are Bradstreet's index of wholesale prices and Persons' price index of business cycles. The constitution of each of these indices was explained in an earlier section.[2] Bradstreet's index, based upon prices of 96 commodities, and Persons' index, based upon prices of 10 commodities, move simultaneously, in general, but the movements of the latter are much more violent. Accordingly, since Persons' object is the measurement of cyclical fluctuations alone, Bradstreet's index is used as a base for the more sensitive 10-commodity index. Deviations are measured from this base precisely as from an ordinary line of trend.

The same method is employed in securing a base line for money rates. A stable series, the yield on 10 prime railroad bonds, is used as base for each of two more sensitive series which move concurrently, the rate on 60–90 day paper and the rate on 4–6 months paper. In each case a slight adjustment was necessary in order that the "crossing points" (i.e., the points at which the line representing each series crossed its base line) might coincide with the "crossing points" of other series used in the general index of business conditions. These adjustments are explained in the references cited.

Deflation as a Step in Analysis

Many series of economic data are expressed in monetary units, in dollars, pounds, or francs, and such series are subject to distortion because of changes in the price level. Thus the value of building and engineering contracts awarded in twenty-seven states in 1913 amounted to 858 millions of dollars; in 1922 the value of contracts awarded in the same territory amounted to 3344 millions of dollars.[3]

[1] *Cf. Review of Economic Statistics*, Harvard Economic Service, April, 1923, 73–74, July, 1923, 192–194.

[2] *Cf.* Chapter VI. [3] Figures compiled by F. W. Dodge & Co.

Was the volume of building in 1922 almost four times that in 1913? It was not. The value of building contracts awarded in any year depends not only upon the actual volume of construction but also upon the costs of building materials and building labor, and these increased materially from 1913 to 1922. If we wish to measure the change in the volume of building alone, these values must be corrected for the increase in building costs between 1913 and 1922. Such a process is termed *deflation*.

The selection of an appropriate deflating index is the central problem in such cases. In the chapter on index numbers, money wages were deflated by a cost of living index to secure a measure of the changes in real wages. We might deflate the value of building contracts by an index of building material prices, at wholesale, in the following manner:

(1)	(2)	(3)	(4)
Year	*Value of Building Contracts (millions of dollars)*	*Index of Building Material Prices*	*Deflated Value of Building Contracts (millions of dollars)*
1913	858	100	858
1922	3344	168	1990

The values in the last column are secured by dividing the actual values in column (2) by the corresponding index numbers in column (3), and multiplying by 100. This has the effect of reducing the actual value to the value in terms of prices during the base period, in the present instance 1913. The actual volume of building is shown by the deflated figures to have increased in 1922 to slightly more than double the 1913 volume.

Better results could be secured in this case by taking account of wage changes, as well as changes in material costs, in deflating. The American Telephone and Telegraph Company uses such a combined index in deflating the value of building contracts awarded. The application of this index is illustrated in the following table:

TABLE 74

Actual and Deflated Values of Building Contracts Awarded

	Value of Building Contracts Awarded (millions of dollars)	*A.T. & T. index of Construction Costs (1914=100)*	*Deflated Value of Building Contracts (millions of dollars)*
1914			
January	51.1	100	51.1
February	39.1	101	38.7
March	58.9	102	57.7
April	79.7	101	78.9
May	72.0	100	72.0
June	81.8	100	81.8
July	72.0	100	72.0
August	77.3	100	77.3
September	47.1	100	47.1
October	53.4	99	53.9
November	45.5	99	46.0
December	42.3	98	43.2
1915			
January	43.3	99	43.7
February	48.8	99	49.3
March	75.6	99	76.4
April	76.5	100	76.5
May	77.1	102	75.6
June	92.2	104	88.7
July	94.7	104	91.1
August	90.4	102	88.6
September	82.0	103	79.6
October	88.6	104	85.2
November	88.0	107	82.2
December	82.9	112	74.0
1916			
January	62.8	118	53.2
February	66.3	121	54.8
March	94.5	124	76.2
April	100.9	127	79.4
May	131.4	127	103.5
June	140.7	128	109.9
July	114.4	125	91.5
August	127.0	122	104.1
September	132.2	123	107.5
October	151.4	125	121.1
November	122.4	128	95.6
December	112.9	136	83.0

TABLE 74 — *Continued*

	Value of Building Contracts Awarded (millions of dollars)	*A. T. & T. index of Construction Costs (1914 = 100)*	*Deflated Value of Building Contracts (millions of dollars)*
1917			
January	90.8	138	65.8
February	95.2	141	67.5
March	132.7	144	92.2
April	148.5	148	100.3
May	157.6	155	101.7
June	206.5	160	129.1
July	159.2	171	93.1
August	165.6	170	97.4
September	122.5	166	73.8
October	154.5	154	100.3
November	94.3	155	60.8
December	90.8	155	58.6
1918			
January	161.6	156	103.6
February	137.3	161	85.3
March	115.3	162	71.2
April	128.9	164	78.6
May	120.4	167	72.1
June	248.2	171	145.1
July	153.0	176	86.9
August	146.4	180	81.3
September	124.5	180	69.2
October	166.1	183	90.8
November	130.3	183	71.2
December	57.3	186	30.8
1919			
January	54.1	186	29.1
February	98.7	186	53.1
March	121.8	186	65.5
April	188.8	178	106.1
May	234.7	178	131.9
June	285.4	180	158.6
July	317.7	183	173.6
August	295.1	194	152.1
September	228.7	199	114.9
October	307.5	201	153.0
November	220.6	207	106.6
December	226.7	221	102.6

TABLE 74 — *Continued*

	Value of Building Contracts Awarded (millions of dollars)	*A. T. & T. index of Construction Costs (1914 = 100)*	*Deflated Value of Building Contracts (millions of dollars)*
1920			
January	226.1	227	99.6
February	205.3	240	85.5
March	302.2	248	121.9
April	304.9	254	120.0
May	264.9	264	100.3
June	260.1	266	97.8
July	204.5	265	77.2
August	202.7	267	75.9
September	182.2	261	69.8
October	178.6	256	69.8
November	132.9	250	53.2
December	100.1	236	42.4
1921			
January	111.6	231	48.3
February	102.4	223	45.9
March	163.8	221	74.1
April	221.3	212	104.4
May	240.9	206	116.9
June	226.2	201	112.5
July	212.2	197	107.7
August	220.4	194	113.6
September	244.9	188	130.3
October	222.4	186	119.6
November	190.8	187	102.0
December	198.3	186	106.6
1922			
January	166.3	185	89.9
February	172.7	185	93.4
March	293.4	187	156.9
April	353.0	186	189.8
May	361.8	188	192.4
June	342.4	195	175.6
July	350.1	199	175.9
August	322.0	197	163.5
September	271.5	201	135.1
October	253.1	201	125.9
November	243.4	197	123.6
December	214.2	196	109.3

TABLE 74 — *Continued*

	Value of Building Contracts Awarded (millions of dollars)	*A. T. & T. index of Construction Costs (1914 = 100)*	*Deflated Value of Building Contracts (millions of dollars)*
1923			
January	218.7	202	108.3
February	229.9	209	110.0
March	338.2	212	159.5
April	362.9	217	167.2
May	373.9	217	172.3
June	323.6	221	146.4
July	274.2	219	125.2
August	253.1	219	115.6
September	253.5	217	116.8
October	319.9	216	148.1
November	289.3	211	137.1
December	267.9	215	124.6

The actual and deflated values are plotted in Fig. 60.

Most value series are affected by price changes, and it is generally advisable to correct for this factor before further analysis is attempted. Each case presents a new problem, for no general deflating index is suitable to all series. The index of wholesale prices compiled by the United States Bureau of Labor Statistics has been used extensively in deflating economic data expressed in dollar values, but this index is not at all appropriate in many of the cases in which it has been employed. It would be absurd, for instance, to deflate money wages by an index of wholesale prices. The deflating index employed should be a measure of price changes as they affect the series being deflated.

An interesting example is afforded by the method adopted by the statistical department of the Federal Reserve Bank of New York in deflating bank clearings outside New York. The deflating index employed was a combination of the following series, each weighted as indicated:

Index of rents	1
Index of wholesale commodity prices	2
Index of wages	$3\frac{1}{2}$
Index of cost of living	$3\frac{1}{2}$

It was found that the index so constructed served quite effectively to eliminate the effect of price and wage changes upon outside bank clearings. The deflated series has been used as a measure of variations in the volume of trade in the United States.

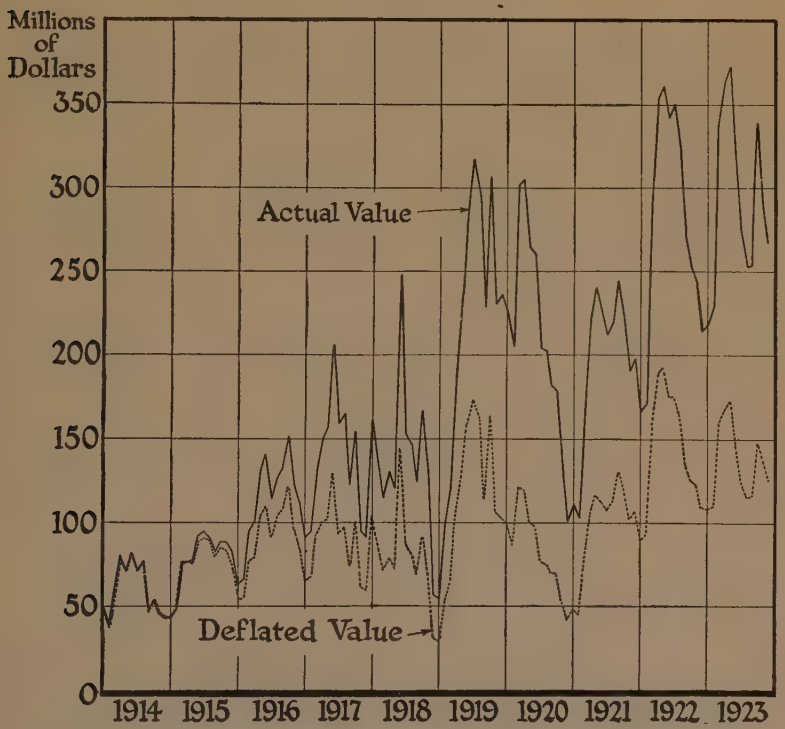


FIG. 60. — Comparison of Actual and Deflated Values of Building Contracts Awarded, by Months, 1914–1923

The deflation of a value series is in general a first step in the study of that series. The way is then open for further analysis by methods explained in the present and succeeding chapters.

REFERENCES

BOWLEY, A. L. *Elements of Statistics* (132–169).

DAVIES, G. R. *Introduction to Economic Statistics* (100–130).

KARSTEN, KARL. *Charts and Graphs* (477–481).

LIPKA, JOSEPH. *Graphical and Mechanical Computation.*

MOORE, H. L. *Forecasting the Yield and the Price of Cotton* (28–37).

PEARL, RAYMOND. *Medical Biometry and Statistics* (332–341).

PERSONS, W. M. *Indices of Business Conditions.* Review of Economic Statistics. Prel. Vol. I, 1919 (1–107).

RUNNING, T. R. *Empirical Formulas.*

STEINMETZ, C. P. *Engineering Mathematics* (209–255).

(References to books dealing with the method of least squares are given at the end of Appendix A.)

CHAPTER VIII

THE ANALYSIS OF TIME SERIES: MEASUREMENT OF SEASONAL AND CYCLICAL FLUCTUATIONS

The measurement of secular trend is but one of the problems connected with the analysis of a series in time. Such series, it has been pointed out, are subject to periodic fluctuations, seasonal and cyclical in character, and these fluctuations are generally more important in their effects upon business than is the long-time trend. Our present concern is with methods of isolating such periodic variations. The series which follows, characterized by extreme seasonal variation, may be used to illustrate methods of measuring such seasonal swings.

TABLE 75

Average Egg Prices Received by Producers in the United States, 1910–1921 [1]

(Prices are given in cents per dozen; averages are based upon prices on the first of each month.)

	1910	*1911*	*1912*	*1913*	*1914*	*1915*	*1916*	*1917*	*1918*	*1919*	*1920*	*1921*
January	30.5	30.4	29.5	26.8	30.7	31.6	30.6	37.7	46.3	57.2	64.8	61.1
February	28.9	22.1	29.1	22.8	28.4	29.2	26.8	35.8	49.4	48.3	56.9	49.6
March	22.9	16.5	24.5	19.4	24.2	21.3	21.2	33.8	40.4	33.1	46.6	29.2
April	18.6	14.9	17.8	16.4	17.6	16.6	17.9	25.9	31.2	34.3	38.8	20.4
May	18.6	14.7	17.1	16.1	16.8	17.1	18.1	30.0	31.0	36.8	37.4	20.2
June	18.3	14.5	16.7	16.9	17.3	16.6	19.0	31.1	29.8	38.6	37.0	19.4
July	18.2	14.2	16.7	17.0	17.6	16.8	19.7	28.3	30.7	36.8	36.7	22.0
August	17.6	15.5	17.4	17.2	18.2	17.0	20.7	29.8	34.4	39.3	40.0	26.6
September	19.4	17.4	19.1	19.5	21.0	18.7	23.3	33.2	36.4	41.0	44.2	30.4
October	22.4	20.0	22.0	23.4	23.5	22.3	28.1	37.4	41.6	44.7	50.1	34.2
November	25.3	23.5	25.9	27.4	25.3	26.3	32.2	39.4	47.2	54.0	56.9	44.2
December	29.0	28.7	29.7	33.0	29.7	30.6	38.1	43.3	55.0	61.9	65.0	51.1
Average	22.5	19.4	22.1	21.3	22.5	22.0	24.6	33.8	39.5	43.8	47.9	34.0

[1] Data compiled by U. S. Department of Agriculture.

Arithmetic Averages of Monthly Items

A comparatively simple and straightforward method of measuring seasonal variations is afforded by the use of arithmetic averages of the actual items, by months. Since the data cover 12 years, there are 12 January items, 12 February items, etc. When the items, by months, are averaged the results shown in column (2) of Table 76 are secured.

Table 76

The Construction of Indices of Seasonal Variation from Arithmetic Averages of Actual Prices

(1)	(2)	(3)	(4)
Month	*Arithmetic averages of monthly egg prices (cents per dozen)*	*Arithmetic averages corrected for secular trend (cents per dozen)*	*Indices of Seasonal Variation*
January.....	39.77	40.73	138.8
February.....	35.61	36.38	123.9
March.....	27.76	28.34	96.5
April.....	22.53	22.92	78.1
May.....	22.82	23.01	78.4
June.....	22.93	22.93	78.1
July.....	22.89	22.70	77.3
August.....	24.47	24.08	82.0
September.....	26.97	26.39	89.9
October.....	30.81	30.04	102.3
November.....	35.63	34.66	118.1
December.....	41.26	40.10	136.6
Average.....	.....	29.357	100.0

The simple arithmetic averages in column (2) give some indication of the extent of seasonal variation in the prices of eggs. From January there is a sharp falling off to April, and then a gradual increase, broken slightly in the month of July, reaching a maximum in December. These figures, reduced to a percentage basis, would serve as a measure of seasonal changes were it not for the effect of secular trend upon the averages. From 1910 to 1920 egg prices increased materially, and this increase, as well as the seasonal varia-

tion, is reflected in the monthly averages. On the average, taking secular trend alone into account, every February exceeds the preceding January, every March exceeds the preceding February, etc., and, accordingly, the monthly averages must be corrected for this factor if seasonal variations alone are to be measured.

When a straight line is fitted to the twelve yearly averages of egg prices from 1910 to 1921, inclusive, it is found that the average annual increase in price during this period was 2.32 cents per dozen. The average monthly increase was one-twelfth of this, or .193 cents. That is, taking no account of seasonal variations, the trend of egg prices was such that every month prices tended to exceed the prices of the preceding month by .193 cents. The average of each month's prices, therefore, is affected by this same amount. Were the force of secular trend alone in operation, the average of the February items would exceed the average of the January items by .193 cents, the March average would exceed the February average by the same amount, etc. The effect of this factor may be easily eliminated by taking any one month as base, and making the necessary corrections in the figures for all other months. With January as base we would subtract from the February average .193 cents, from the March average .386 cents, from the April average .579 cents, etc. With December as base there should be added to the November average .193 cents, to the October average .386 cents, etc. It is advisable for the present purpose to take a base near the middle of the year. Thus with June as base we subtract .193 cents from the July average, .386 cents from the August average, etc., while we add .193 cents to the May average, .386 cents to the April average, etc. When these corrections are made the averages given in column (3) of Table 76 are secured.

These figures represent the effect of seasonal fluctuations alone, during the twelve years included in the study. For convenience in use it is advisable to reduce them to a

percentage form, with the average monthly figure (29.357) as the base. These percentages are given in column (4) of Table 76. The use of such indices as these in the analysis of time series is demonstrated below.

The Method of Link Relatives

Another method of measuring seasonal variation has been developed by Warren M. Persons. This method has been used in several important investigations, notably in the construction of indices of business conditions by Persons. The first step in this method involves the computation of *link relatives*, by expressing the figure for each month of the period covered as a percentage of the figure for the preceding month. Thus January, 1910, is expressed as a percentage of December, 1909; February, 1910, as a percentage of January, 1910, etc. The following are the link relatives thus computed from the data of Table 75.

Table 77

Link Relatives of Monthly Egg Prices

Month	*1910*	*1911*	*1912*	*1913*	*1914*	*1915*	*1916*	*1917*	*1918*	*1919*	*1920*	*1921*
January	107.4	104.8	102.8	90.2	93.0	106.4	100.0	99.0	106.9	104.0	104.7	94.0
February	94.8	72.7	98.6	85.1	92.5	92.4	87.6	95.0	106.7	84.4	87.8	81.2
March	79.2	74.7	84.2	85.1	85.2	72.9	79.1	94.4	81.8	68.5	81.9	58.9
April	81.2	90.3	72.7	84.5	72.7	77.9	84.1	76.6	77.2	103.6	83.3	69.9
May	100.0	98.7	96.1	98.2	95.5	103.0	101.1	115.8	99.4	107.3	96.4	99.0
June	98.4	98.6	97.7	105.0	103.0	97.1	105.0	103.7	96.1	104.9	98.9	96.0
July	99.5	97.9	100.0	101.0	101.7	101.2	103.7	91.0	103.0	95.3	99.2	113.4
August	96.7	109.2	104.2	101.2	103.4	101.2	105.1	105.3	112.1	106.8	109.0	120.9
September	110.2	112.3	109.8	113.4	115.4	110.0	112.6	111.4	105.8	104.3	110.5	114.3
October	115.5	114.9	115.2	120.0	111.9	119.3	120.6	112.6	114.3	109.0	113.3	112.5
November	112.9	117.5	117.7	117.1	107.7	117.9	114.6	105.3	113.5	120.8	113.6	129.2
December	114.6	122.1	114.7	120.4	117.4	116.3	118.3	109.9	116.5	114.6	114.2	115.6

It remains to determine the average relationship of January prices to December prices, of February prices to January prices, etc. In the preceding example this was done by getting arithmetic averages of each month's prices and reducing these to percentages. In employing Persons'

method the median link relatives are found. That is, all the January link relatives are grouped in order of magnitude, and the median value determined, the same being done for each of the other months. The median link relatives thus secured are given in column (2) of Table 78.

TABLE 78

The Construction of Indices of Seasonal Variation by the Method of Link Relatives

(1)	(2)	(3)	(4)	(5)
Month	*Median Link Relatives of Monthly Egg Prices*	*Chain Relatives*	*Corrected Chain Relatives*	*Indices of Seasonal Variation*
January.....	103.4	100.0	100.0	139.5
February.....	90.1	90.1	89.7	125.1
March.....	80.5	72.5	71.9	100.3
April.....	79.5	57.4	56.7	79.1
May.....	99.2	56.9	56.0	78.1
June.....	98.7	56.2	55.1	76.9
July.....	100.5	56.5	55.2	77.0
August.....	105.2	59.4	57.8	80.6
September....	111.0	65.9	63.9	89.1
October.....	114.6	75.5	72.9	101.7
November....	115.8	87.4	84.0	117.2
December....	116.0	101.4	97.1	135.4
Average....	.....	.....	71.69	100.0

These median link relatives measure the monthly variations in terms of a shifting base, each month being the base for the one which follows. It is next necessary to express the monthly variations as percentages of a constant base. To this end *chain relatives* are constructed, with January as base. The January figure is thus 100; the February figure is 90.1, which corresponds to the median link relative, since all the February link relatives are computed on January as base. But the March median link relative is 80.5 with February as a base; the March chain relative, with January as base, will be 80.5×90.1 or 72.5. Similarly the chain relative for each month is computed by multi-

plying the median link relative of that month by the chain relative of the month preceding. The figures shown in column (3) of Table 78 are secured.

If there were no upward or downward trend in the data the product secured by multiplying the December chain relative by the January median link relative would be 100. Actually the product of 101.4 and 103.4 (the latter being the January median link relative) is 104.8. The existence of a constant trend, affecting each of the twelve median link relatives, is responsible for the difference. This, it should be noted, is the result of a cumulative error, compounded twelve times when the median link relatives are multiplied by the successive chain relatives. The case is identical with the compounding of a sum of money at a fixed rate of interest. The error in the present case (the rate of monthly increase due to secular trend) corresponds to the rate of interest. The formula for the growth of a fixed sum (P_0) compounded at a constant rate of interest over n years is as follows (P_n representing the principal at the end of n years):

$$P_n = P_0(1 + r)^n.$$

The rate of growth is thus represented by

$$r = \sqrt[n]{\frac{P_n}{P_0}} - 1.$$

The values secured in the present case may be inserted in these formulas:

$$104.8 = 100(1 + r)^{12}$$

$$r = \sqrt[12]{\frac{104.8}{100}} - 1$$

$$= 1.004 - 1$$

$$= .004.$$

The increase due to secular trend is at the rate of .004 per month. To eliminate the effect of this factor the February chain relative must be divided by 1.004, that is,

by $(1 + r)$, the March chain relative by 1.008, $(1 + r)^2$, etc., the December chain relative being divided by 1.044, or $(1 + r)^{11}$. In this way the effect of long-term trend is eliminated from the chain relatives which are to be used as measures of seasonal variation. These corrected chain relatives are given in column (4) of Table 78.

It is desirable that an index of seasonal variation be based upon an average monthly figure, rather than upon the January figure. With the average as base the total of the twelve monthly percentages will be 1200, and the interpretation and use of each monthly figure is somewhat simpler. When the monthly figures in the preceding table are thus expressed as percentages of their average (71.69) the indices in column (5) of Table 78 are secured.

The Use of Moving Averages

Another method of measuring seasonal variations is offered by the device of moving averages. Since these fluctuations take place within a constant period of twelve months, a moving average may be used with more confidence than when a cycle of varying length is involved. The magnitude of the fluctuations (the *amplitude* of the seasonal swings) will not ordinarily be constant, hence the line marked out by the moving averages will not be completely free of seasonal influences. The relation of the actual monthly items to the moving averages may be averaged, however, and the indices of seasonal variation based upon these averages.

It is essential, of course, that the moving average, centered, fall at the same date as the original figure with which it is to be compared. This involves a second process of averaging. For example, the egg prices given are as of the first of each month. The average of the twelve monthly items for 1910, when centered, falls on June 15th. The average of the items from February, 1910, through January,

1911, centered, falls on July 15th. To secure a figure comparable with the July 1st price, these two must be averaged. By this process of computing a two-month moving average from the twelve-month moving average comparability with the original figures is secured. The following table presents the averages obtained in this way:

TABLE 79

Moving Averages of Egg Prices, 1910–1921

(12-month moving average, centered, adjusted by 2-month moving average, centered)
(Prices are given in cents per dozen)

Month	*1910*	*1911*	*1912*	*1913*	*1914*	*1915*	*1916*	*1917*	*1918*	*1919*	*1920*	*1921*
January	.....	20.25	21.26	20.78	22.74	22.29	22.24	30.06	36.72	41.39	46.60	40.47
February	.....	19.99	21.45	20.79	22.80	22.20	22.51	30.80	37.01	41.86	46.63	39.30
March	.....	19.82	21.60	20.79	22.91	22.06	22.86	31.59	37.33	42.25	46.80	38.16
April	.....	19.64	21.75	20.87	22.97	21.91	23.29	32.39	37.64	42.57	47.15	36.93
May	.....	19.46	21.94	20.99	22.89	21.90	23.77	33.08	38.14	42.98	47.50	35.74
June	.....	19.38	22.08	21.19	22.66	21.98	24.33	33.59	38.96	43.56	47.75	34.63
July	22.47	19.33	22.00	21.49	22.57	21.98	24.94	34.17	39.90	44.17	47.72	.....
August	22.18	19.58	21.63	21.88	22.64	21.84	25.60	35.10	40.31	44.84	47.26	.....
September	21.63	20.20	21.16	22.32	22.55	21.74	26.51	35.93	39.96	45.76	46.24	.....
October	21.21	20.66	20.89	22.57	22.39	21.78	27.36	36.43	39.79	46.50	44.74	.....
November	20.89	20.88	20.79	22.65	22.36	21.88	28.20	36.69	40.17	46.71	43.26	.....
December	20.57	21.07	20.76	22.69	22.35	22.02	29.20	36.68	40.77	46.67	41.81	.....

The actual prices are now expressed as percentages of the corresponding moving averages. These percentages are given below:

TABLE 80

Percentage Relation of Actual Egg Prices to 12-Month Moving Averages

Month	*1910*	*1911*	*1912*	*1913*	*1914*	*1915*	*1916*	*1917*	*1918*	*1919*	*1920*	*1921*
January	.....	150.1	138.8	129.0	135.0	141.8	137.6	125.4	126.1	138.2	139.1	151.0
February	.....	110.6	135.7	109.7	124.6	131.5	119.1	116.2	133.5	115.4	122.0	126.2
March	.....	83.2	113.4	93.3	105.6	96.6	92.7	107.0	108.2	78.3	99.6	76.5
April	.....	75.9	81.8	78.6	76.6	75.8	76.9	80.0	82.9	80.6	82.8	55.2
May	.....	75.5	77.9	76.7	73.4	78.1	76.1	90.7	81.3	85.6	78.7	56.5
June	.....	74.8	75.6	79.8	76.3	75.5	78.1	92.6	76.5	88.6	77.5	56.0
July	81.0	73.5	75.9	79.1	78.0	76.4	79.0	82.8	76.9	83.3	76.9	.....
August	79.4	79.2	80.4	78.6	80.4	77.8	80.9	84.9	85.3	87.6	84.6	.....
September	89.7	86.1	90.3	87.4	93.1	86.0	87.9	92.4	91.1	89.6	95.6	.....
October	105.6	96.8	105.3	103.7	105.0	102.4	102.7	102.7	104.5	96.1	112.0	.....
November	121.1	112.5	124.6	121.0	113.1	120.2	114.2	107.4	117.5	115.6	131.5	.....
December	141.0	136.2	143.1	145.4	132.9	139.0	130.5	118.1	134.9	132.6	155.5	.....

These percentages show considerable variation from year to year in the relation of the figures for a given month to

the moving average. Thus the January figures, while always above the average, vary from 125.4 per cent to 151 per cent of the average. The eleven percentages secured for each month must be averaged to obtain the index desired. Either the arithmetic average or the median may be employed for this purpose. The results secured by applying the two methods are shown in the following table. In columns (2) and (3) the actual arithmetic means and medians are given. These indices may be rendered more serviceable by adjustments which make the average in each case equal to 100. These adjustments have been made in computing the index numbers appearing in columns (4) and (5).

TABLE 81

Indices of Seasonal Variation in Egg Prices Computed from Moving Averages

(1)	(2)	(3)	(4)	(5)
Month	*Arithmetic Means (Unadjusted)*	*Medians (Unadjusted)*	*Arithmetic Means (Adjusted)*	*Medians (Adjusted)*
January.....	137.4	138.2	137.9	138.7
February.....	122.2	122.0	122.6	122.5
March.....	95.9	96.6	96.2	97.0
April.....	77.0	78.6	77.2	78.9
May.....	77.3	77.9	77.6	78.2
June.....	77.4	76.5	77.7	76.8
July.....	78.4	78.0	78.7	78.3
August.....	81.7	80.4	82.0	80.7
September....	89.9	89.7	90.2	90.1
October.....	103.3	103.7	103.7	104.1
November....	118.1	117.5	118.5	118.0
December....	137.2	136.2	137.7	136.7
Average....	99.65	99.608	100.0	100.0

THE COMPUTATION OF INDEX NUMBERS OF SEASONAL VARIATION BY AVERAGING RATIOS TO TREND

A further method of securing seasonal indices, which has certain distinctive advantages, has recently been devoloped.[1]

[1] The essentials of this method were worked out independently by Helen D. Falkner ("The Measurement of Seasonal Variation," *Journal of the American*

In the application of this method, a suitable line of trend, linear or non-linear, is fitted to the data, the actual monthly items are expressed as percentages of the corresponding trend figures, and then, for each month, an average of the percentage ratios of the actual to the trend values is secured. This procedure is identical with that described in connection with the use of moving averages, except that the actual values may be expressed as percentages of normal values derived from any function employed to represent trend. In the selection of an average value for each month, use may be made of a multiple frequency table in obtaining an understanding of the nature of the actual seasonal movement. With the help of such a table the existence of a definite seasonal movement may be verified and the type of average to be used in securing a typical value for each month may be determined. We shall apply this method to the data employed in the preceding examples.

A straight line, fitted to annual average egg prices from 1910 to 1921, as given in Table 75, is described by the equation

$$y = 29.45 + 2.32x$$

with origin at Jan. 1, 1916. Normal values for each month may be computed readily, by methods previously described. Expressing the actual values as percentages of the trend values we secure for each month twelve such percentage ratios (since the data cover twelve years). The twelve January percentages vary from 103.5 to 194.8, the twelve May percentages from 48.2 to 113.7, etc. The multiple frequency table which appears in Fig. 61 is constructed by classifying, in the form of a frequency distribution, the items for each month.

The presence of a distinct seasonal variation is demonstrated by this table. For the winter months the per-

---

Statistical Association, June, 1924, 167–179) and Lincoln W. Hall ("Seasonal Variation as a Relative of Secular Trend," *Journal of the American Statistical Association*, June, 1924, 156–166).

centages are consistently higher than for the spring and summer months, and we may conclude that there is a pronounced seasonal variation capable of representation by index numbers.

This table facilitates the selection of a type of average for the measurement of the seasonal movements. The median is likely to be unrepresentative, and subject to material

	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
170 and Over	/	/										
165 – 169.9	//											
160 – 164.9												//
155 – 159.9	/											/
150 – 154.9												/
145 – 149.9	//	/										
140 – 144.9		//	/								//	/
135 – 139.9	/										/	
130 – 134.9		/									/	//
125 – 129.9										/		/
120 – 124.9	/	//								/		/
115 – 119.9	///		///							//	///	/
110 – 114.9		//		/	/	/			/	/	/	
105 – 109.9		/					/		//			/
100 – 104.9	/	/	/			/		//	/	/	//	/
95 – 99.9			/	/	/		/	//	/	//		
90 – 94.9		/		/	//	//	/				//	
85 – 89.9			//	//	/		//	/	//	//		
80 – 84.9			/	/	/	/		//	/			
75 – 79.9			/	/	/	//	/		/	//		
70 – 74.9			//	/		/	//	/	//			
65 – 69.9				/	//	/	/	//	/			
60 – 64.9					/	/	/	/				
55 – 59.9				//	/	/	/	/				
50 – 54.9							/					
Under 50				/	/	/						

FIG. 61. — Frequency Distributions of Ratios of Monthly Egg Prices to the Corresponding Ordinates of Secular Trend. (Based on data for the period 1910–1921)

change in value by the addition or withdrawal of one or two entries, unless there is a definite concentration in the monthly frequency distributions. The arithmetic mean of all the items, on the other hand, will be unduly affected

by exceptional cases. An alternative method is provided by the possibility of taking the arithmetic mean of the central items for each month. If an inspection of the multiple frequency table does not lead to an immediate decision as to which is the best type of average to employ in a given case, several index numbers may be worked out for each month, and a decision reached after a comparison of the results. From the percentage ratios of actual to trend values, as computed from the data of Table 75, six different averages have been worked out for each month. The simple averages constitute the unadjusted index numbers given below. Correcting these so that the average of each group is equal to 100, we secure the adjusted index numbers presented in the same table. (These averages have been derived, not from the frequency distributions shown in Fig. 63, but from the individual percentages of actual to trend values.)

TABLE 82

Unadjusted and Adjusted Index Numbers of Seasonal Variation in Egg Prices

(Based upon Percentage Ratios of Actual Values to Linear Trend Values.)

Month	*Unadjusted Index Numbers (based upon given number of central items)*						*Adjusted Index Numbers (based upon given number of central items)*					
	2	4	6	8	10	12	2	4	6	8	10	12
January	140.5	138.2	138.0	139.2	139.8	141.4	142.1	139.4	138.4	139.1	139.6	140.5
February	120.8	121.6	123.4	123.8	123.6	125.7	122.2	122.6	123.8	123.7	123.5	124.9
March	92.8	95.0	96.9	97.1	96.7	98.3	93.9	95.8	97.2	97.1	96.6	97.7
April	80.0	79.0	79.1	78.1	78.3	78.9	80.9	79.7	79.3	78.1	78.2	78.4
May	79.8	79.3	79.1	78.8	78.8	79.2	80.7	80.0	79.3	78.8	78.7	78.7
June	77.6	77.9	78.6	78.6	79.0	78.9	78.5	78.6	78.8	78.6	78.9	78.4
July	76.1	77.3	77.3	77.5	77.7	78.1	77.0	77.9	77.5	77.5	77.6	77.6
August	80.3	80.5	81.4	81.8	82.1	82.0	81.2	81.2	81.6	81.8	82.0	81.5
September	88.3	89.2	89.7	90.1	90.2	90.0	89.3	89.9	90.0	90.1	90.1	89.5
October	100.7	102.2	102.4	102.4	102.3	102.4	101.8	103.1	102.7	102.4	102.2	101.8
November	116.6	115.7	116.0	117.5	117.3	117.3	117.9	116.7	116.3	117.5	117.2	116.6
December	133.0	134.0	134.7	135.6	135.5	135.2	134.5	135.1	135.1	135.6	135.4	134.4

The unadjusted index numbers in the first column of figures are the arithmetic averages of the two middle items

in each monthly frequency distribution; the figures in the next column are the arithmetic averages of the four central items, etc.

In the present instance the seasonal variation is so pronounced and regular, year after year, that no striking differences appear in the results secured by averaging varying numbers of items. The index numbers based upon averages of the four central items for each of the twelve months are perhaps the best to employ in this case. In general, an average of three, four or five central values is more likely to be stable and representative than either the median or an average of all the items for each month. The greater the concentration in the monthly frequency tables, the smaller the number of items upon which the index numbers may be based.

In the following table four sets of index numbers of seasonal variation in egg prices, secured by the different methods illustrated above, have been brought together for comparison.

TABLE 83

Indices of Seasonal Variation in Egg Prices

Month	*Computed from Moving Average Medians*	(Adjusted) *Computed from Arithmetic Means of Monthly Prices*	*Computed from Link Relatives*	*Computed from Percentage Ratios to Trend (Average of 4)*
January.....	138.7	138.8	139.5	139.4
February.....	122.5	123.9	125.1	122.6
March.....	97.0	96.5	100.3	95.8
April.....	78.9	78.1	79.1	79.7
May.....	78.2	78.4	78.1	80.0
June.....	76.8	78.1	76.9	78.6
July.....	78.3	77.3	77.0	77.9
August.....	80.7	82.0	80.6	81.2
September....	90.1	89.9	89.1	89.9
October.....	104.1	102.3	101.7	103.1
November....	118.0	118.1	117.2	116.7
December....	136.7	136.6	135.4	135.1
Average....	100.0	100.0	100.0	100.0

Comparison of Indices of Seasonal Variation

These four indices show a fairly close agreement as to the seasonal variations in egg prices. Discrepancies of some importance occur in the months of February and March, the difference between the two extreme figures amounting to more than four per cent in the latter month. There is no absolute criterion by means of which the best method of measuring seasonal variation may be determined but certain general conclusions may be suggested. The arithmetic mean of actual prices, while simple in calculation, should not be applied if extreme variations are present, or if the data change materially in value during the period covered. The use of link relatives involves a considerable amount of routine calculation, and in the analysis of homogeneous and fairly regular series this method does not appear to be superior to the simpler methods.[1] The averaging of percentage ratios to moving averages or to normal values determined by fitting appropriate lines of trend is probably the most generally applicable and most useful method. The arithmetic operations are simple, particularly when normal values are based upon mathematical lines of trend, and there is sound theoretical justification for the averaging of ratios rather than absolute values.

[1] W. M. Persons points out the following advantages of the method of link relatives. (The first two of these, it may be noted, are shared by the methods which employ ratios to moving averages and to mathematical lines of trend.)

a. The frequency distributions of link relatives enable one to judge the degree of regularity of month-to-month or seasonal changes.

b. The use of the median (or average of central items) is a device by which the influence of large non-seasonal variations may be greatly moderated.

c. It is possible to use non-homogeneous statistical series in the measurement of seasonal variation. For instance, link relatives for bank clearings of 50 representative cities for one interval may be combined with link relatives for the clearings of 100 representative cities for another interval.

— *Journal of the American Statistical Association*, June, 1923, 717.

The Determination of Monthly Normal Values

Before proceeding to a problem which involves the application of corrections for both trend and seasonal variations, the steps to be taken in making the transition from an annual to a monthly basis, in the computation of normal or trend values, should be outlined. Only simple arithmetic calculations are involved.

The following figures, arbitrarily selected, may be used to exemplify the procedure.

Year	*Normal Annual Production, Commodity X*
1910	186
1911	330
1912	474
1913	618

The data are plotted in Fig. 62, A.

These figures represent the normal production during four successive years. The line of trend in this case is described by the equation

$$y = 186 + 144x$$

with origin at 1910. Since each annual figure is a total, it is to be considered as centered at the middle of the period covered (i.e., July 1st) and the line is drawn through the points so located. This line has a slope of 144, the annual increase in production during this period.

We may consider the reduction to a monthly basis to be made in two steps. First, we secure an equation in which the annual production is expressed in terms of monthly averages. This is done by dividing the values 186 and 144, as given in the preceding equation, by 12, securing

$$y = 15.5 + 12x$$

with origin at the middle of 1910. This equation is plotted in Fig. 62, B. Each of the narrow columns in this diagram represents the production during a month centering at the

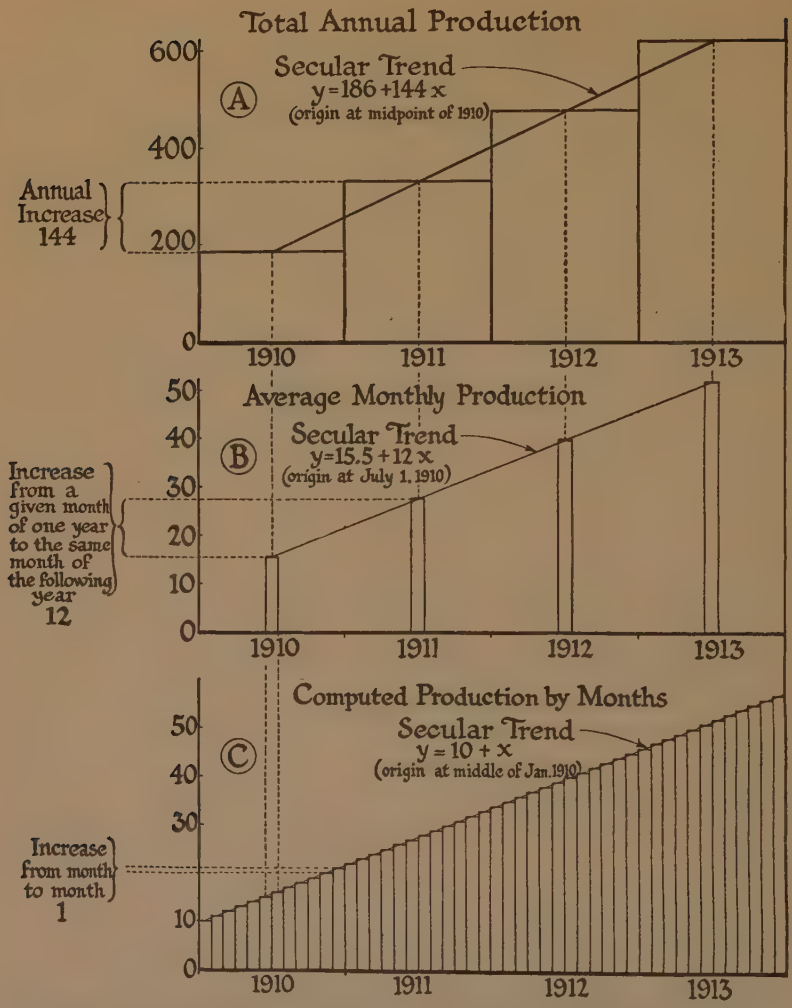


FIG. 62. — Illustrating the Method of Determining Monthly Trend Values (The x-unit is one year in (A) and (B), one month in (C)

middle of the year in question (i.e., July 1st). The slope of the line in this figure is 12, which represents the increase in production during a given month in one year as compared with the same month in the year preceding. Thus normal production for the month centering at July 1, 1910 is 15.5;

normal production for the month centering at July 1, 1911, is 27.5, etc. The x-unit in this equation, it should be noted, is one year.

The next step is the computation of the monthly normal or trend figures. If the increase in normal monthly production between July 1, 1910, and July 1, 1911, is 12, the increase from July 1, 1910, to August 1, 1910, is 1. Thus, for convenience, we may change the abscissal unit from one year to one month. With this unit, the equation would be

$$y = 15.5 + x$$

with origin at July 1, 1910. This is not a convenient origin, since actual monthly data would be considered as centered at the middle of each month. If the ordinate at July 1, 1910, is 15.5, and the monthly increment is 1, the ordinate at January 15, 1910, would be 10. Shifting the origin to January 15, 1910, accordingly, we have the equation

$$y = 10 + x$$

From this equation the normal production for any month may be readily determined. Figure 62, C, illustrates the relation between normal production in successive months during the period 1910–1913.

In Fig. 63 the relations between total annual production (normal), average monthly production (normal) and monthly normals for the year 1910 are brought out in greater detail.[1] The particular point which is emphasized in this chart is the necessity for a correction of one half month in deriving the monthly normals from the average monthly figures.

In series such as that illustrated above the monthly increment due to trend is equal to the annual increment divided by 144. This is true in all cases in which the annual figure is the *sum* of the twelve monthly figures, as with all data relating to production. In price series or series of any other type in which the annual figure is the *average* of the twelve monthly figures, the monthly increment is equal to

[1] Figs. 62 and 63 were devised and prepared by D. H. Davenport.

the annual increment divided by 12. The first of the two steps illustrated above is unnecessary in such cases.

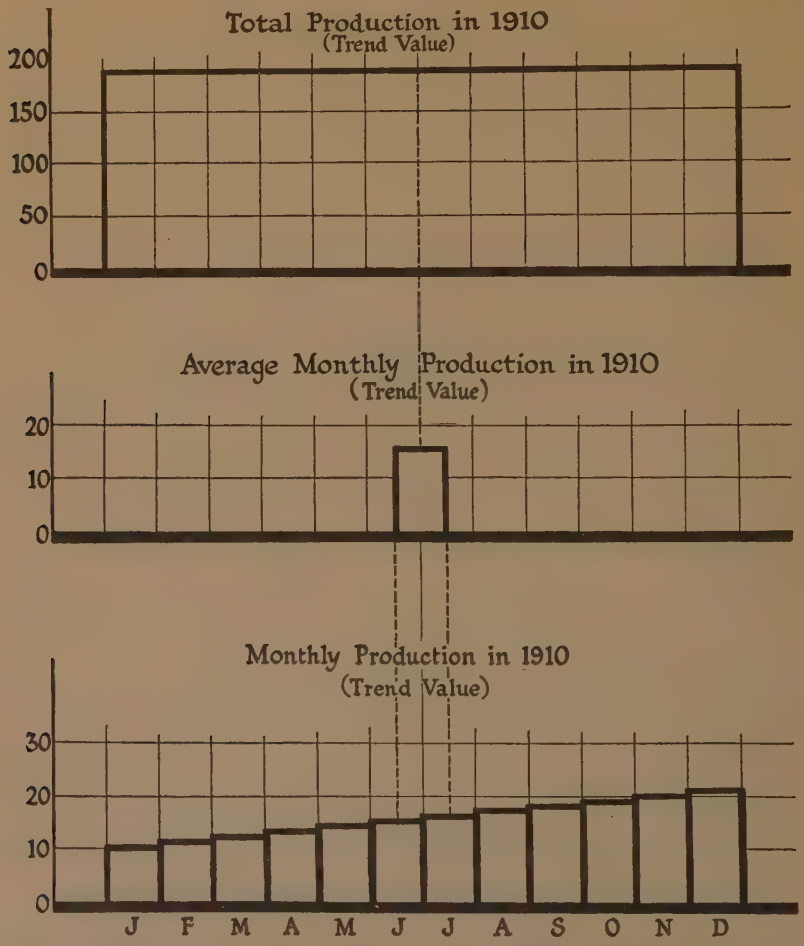


FIG. 63. — Showing the Relation between Trend Values of Total Annual Production, Average Monthly Production and Production by Months

THE MEASUREMENT OF CYCLICAL FLUCTUATIONS

There remains the task of combining the corrections for secular trend and seasonal variation in order to secure measures of cyclical changes in a given series. Major

interest in most economic studies attaches to these cyclical changes, and the measurement of such changes is usually the central problem in the analysis of time series. To illustrate the procedure we shall analyze the figures for bituminous coal production in the United States, from 1901 to 1923.

An appropriate line of trend must first be fitted to the annual production figures for this period. An equation of the type

$$\log y = a + bx + cx^2$$

is selected, and the values of the constants are determined by the method of least squares. The equation to the curve which describes the trend is

$$\log y = 2.646052 + .014226x - .0010246x^2$$

From this equation the normal production for the several years may be computed. These are given in the following table, in comparison with the actual figures.

TABLE 84

Bituminous Coal Production in the United States
Comparison of Actual and Normal Values, 1901–1923

(1)	(2)	(3)	(4)	(5)
Year	*Actual Production (in millions of net tons)*	*Logarithm of Trend*	*Normal Production (in millions of net tons)*	*Actual Production as Percentage of Normal*
1901	225.83	2.365589	232.05	97.3
1902	260.22	2.401332	251.96	103.2
1903	282.75	2.435025	272.29	103.8
1904	278.66	2.466670	292.87	95.1
1905	315.06	2.496265	313.52	100.5
1906	342.87	2.523810	334.05	102.6
1907	394.76	2.549307	354.25	111.4
1908	332.57	2.572754	373.90	88.9
1909	379.74	2.594153	392.78	96.7
1910	417.11	2.613502	410.68	101.6
1911	405.91	2.630801	427.36	95.0
1912	450.10	2.646052	442.64	101.7
1913	478.44	2.659253	456.30	104.9
1914	422.70	2.670406	468.17	90.3
1915	442.62	2.679509	478.09	92.6
1916	502.52	2.686562	485.92	103.4

TABLE 84 (*Continued*)

(1) Year	(2) Actual Production (in millions of net tons)	(3) Logarithm of Trend	(4) Normal Production (in millions of net tons)	(5) Actual Production as Percentage of Normal
1917	551.79	2.691567	491.55	112.3
1918	579.38	2.694522	494.91	117.1
1919	465.86	2.695429	495.94	93.9
1920	568.67	2.694286	494.64	115.0
1921	415.92	2.691093	491.01	84.8
1922	404.51	2.685852	485.12	83.4
1923	545.30	2.678561	477.05	114.3

From this table the cycles in bituminous coal production, as shown by annual production figures, may be determined directly. Our present problem, however, is the measurement of cyclical changes in coal production, as reflected in monthly figures.

The first step is the determination of the ordinate of the line of trend for each month of the period covered. This ordinate measures the "normal" price for that month. The line of trend has been fitted to annual data, and it is necessary to effect the transition to a monthly basis.

Each of the figures entering into this series is to be considered as representing the middle of the period covered. That is, the total production for the year 1913 is to be considered as centered at July 1, 1913; the figure for production in January, 1913, is to be taken as of January 15, 1913.

The trend value for 1913 is 456.30 millions of tons. This is taken as normal for the year centering at July 1, 1913. Dividing this by 12, we have 38.025 millions of tons as the normal production for the month centering at July 1, 1913. As this figure stands, it cannot be compared with either the June or the July figure for actual production, for there is a discrepancy of 15 days. This may be readily adjusted. The normal value for the year 1914, divided by 12, gives 39.014 millions of tons as normal for the month centering at July 1, 1914. The difference between 39.014 and 38.025, or .989, represents the difference in average monthly pro-

duction, due to trend, between a given month of one year and the same month of the preceding year. The increment from one month to the next is equal to .989 divided by 12, or .0824. This represents the increase from July 1, 1913, to August 1, 1913. The increase from July 1, 1913, to July 15, 1913, is one half of this, or .0412. It has been shown that normal for the month centering at July 1, 1913, is 38.025 millions of tons. Normal for the month centering at July 15, 1913, is 38.025 + .0412, or 38.0662 millions of tons. Normal for the month centering at August 15, 1913, is 38.0662 + .0824, or 38.1486. The trend value for the month of September is 38.1486 + .0824, or 38.2310. The normal production for each of the months between July, 1913, and June, 1914, may be computed in a similar fashion.

Were the trend of the series linear throughout, monthly normals for the entire period could be computed in this way by adding a constant increment to each monthly normal to secure the normal value for the following month. But in this series the rate of normal growth changes from year to year,[1] and the increment of normal value changes correspondingly. The table on page 336 presents the normal or trend figure for July of each year (centering at the 15th of each July), with the monthly increment of normal production for the various years, for the period 1912–1922, inclusive.

This change in the rate of growth must be taken account of in computing the monthly normals. The process which has been explained in detail for the period July, 1913, to June, 1914, must be repeated for each twelve-month period.

It remains to compute the indices of seasonal variation. This may be done by any of the methods which have been described above. In the present instance these indices were computed by the method of moving averages, employing

[1] Strictly, the rate of normal growth changes from month to month. Interpolation on the assumption of a linear trend for each twelve-month period involves no appreciable error, however, and is much easier than with a more complicated equation.

TABLE 85

Changes in the Rate of Normal Growth, Bituminous Coal Production, 1912–1922

Date	*Normal Production (in millions of net tons)*	*Monthly Increment of Normal Production for twelve months following the given date (in millions of net tons)*
July, 1912	36.9341	.0948
July, 1913	38.0662	.0824
July, 1914	39.0490	.0690
July, 1915	39.8680	.0543
July, 1916	40.5120	.0391
July, 1917	40.9742	.0233
July, 1918	41.2461	.0071
July, 1919	41.3238	− .0090
July, 1920	41.2074	− .0252
July, 1921	40.8970	− .0409
July, 1922	40.3987	− .0560

monthly production figures for the period 1913–1921.[1] The values are shown in Table 86.

The first step in the computation of cycles in bituminous coal prices is the reduction of the actual monthly figures to percentages of the monthly normal or trend figures. The results are shown in column (4) of Table 86, which illustrates the process for two of the years covered in the study. These percentages are to be corrected for seasonal fluctuations, the indices of which are given in column (5). It is obvious that if the percentages in column (4) are less than the corresponding percentages in column (5), the deviations of the actual figures from normal are negative, and if the reverse is true they are positive. The deviations in column (6) are secured by subtracting the indices of seasonal variation from the percentages in column (4). The items in column (6) are thus measures of the cyclical movements in bituminous coal production, the effects of both secular trend and seasonal fluctuations having been eliminated. The

[1] The indices of seasonal variation in bituminous coal production which have been employed in this problem were computed by the Statistical Department of the Federal Reserve Bank of New York.

effects of accidental and irregular movements have not, of course, been removed. The indices in column (6) are taken to measure cyclical changes on the assumption that such accidental changes are mutually offset in the long run. (In the case of coal, this assumption is not always justified.)

TABLE 86

Computation of Cycles in Bituminous Coal Production

1913–1914

Year	(1) Month	(2) Actual Production (in millions of net tons)	(3) Normal Production (in millions of net tons)	(4) Actual as Percentage of Normal (2) ÷ (3)	(5) Indices of Seasonal Variation	(6) Percentage Deviations from normal corrected for Seasonal Variation (4) − (5)
	January	42.274	37.503	112.7	107	+5.7
	February	37.057	37.598	98.6	92	+6.6
	March	37.536	37.693	99.6	102	−2.4
	April	34.169	37.788	90.4	83	+7.4
	May	37.205	37.882	98.2	92	+6.2
1913	June	37.405	37.977	98.5	95	+3.5
	July	38.858	38.066	102.1	97	+5.1
	August	41.590	38.149	109.0	106	+3.0
	September	41.424	38.231	108.4	106	+2.4
	October	46.164	38.313	120.5	113	+7.5
	November	43.233	38.396	112.6	104	+8.6
	December	41.519	38.478	107.9	103	+4.9
	January	40.191	38.561	104.2	107	−2.8
	February	35.472	38.643	91.9	92	−.1
	March	45.455	38.725	117.4	102	+15.4
	April	23.609	38.808	60.8	83	−22.2
	May	28.551	38.890	73.4	92	−18.6
1914	June	31.412	38.973	80.6	95	−14.4
	July	34.305	39.049	87.9	97	−9.1
	August	37.751	39.118	96.5	106	−9.5
	September	39.019	39.187	99.6	106	−6.4
	October	37.685	39.256	96.0	113	−17.0
	November	33.392	39.325	84.9	104	−19.1
	December	35.862	39.394	91.0	103	−12.0

The calculations are shown in detail for only two years. When the same process is repeated for later years, the results presented in the following table are secured. The actual monthly data for the period 1913–1923 are shown, together with the percentage deviations from normal corrected for seasonal variation.

TABLE 87

Bituminous Coal Production, 1913–1923

Actual Monthly Production and Corrected Deviations from Normal

	1913		1914		1915	
	Actual Production (in millions of net tons)	*Corrected Deviations from normal (percentage)*	*Actual Production (in millions of net tons)*	*Corrected Deviations from normal (percentage)*	*Actual Production (in millions of net tons)*	*Corrected Deviations from normal (percentage)*
January	42.274	+ 5.7	40.191	− 2.8	37.194	− 12.7
February	37.057	+ 6.6	35.472	− .1	29.321	− 17.8
March	37.536	− 2.4	45.455	+ 15.4	31.801	− 21.7
April	34.169	+ 7.4	23.609	− 22.2	29.968	− 7.5
May	37.205	+ 6.2	28.551	− 18.6	30.938	− 14.1
June	37.405	+ 3.5	31.412	− 14.4	33.957	− 9.7
July	38.858	+ 5.1	34.305	− 9.1	35.573	− 7.8
August	41.590	+ 3.0	37.751	− 9.5	38.161	− 10.4
September	41.424	+ 2.4	39.019	− 6.4	40.964	− 3.5
October	46.164	+ 7.5	37.685	− 17.0	44.198	− 2.6
November	43.283	+ 8.6	33.392	− 19.1	44.737	+ 7.6
December	41.519	+ 4.9	35.862	− 12.0	45.814	+ 11.1
	1916		1917		1918	
January	46.593	+ 8.9	47.969	+ 10.7	42.227	− 4.3
February	45.187	+ 20.3	41.353	+ 9.4	43.777	+ 14.4
March	43.829	+ 6.8	47.869	+ 15.3	48.113	+ 14.9
April	33.628	+ .3	41.854	+ 19.4	46.041	+ 28.8
May	38.804	+ 4.0	47.086	+ 23.1	50.443	+ 30.4
June	37.742	− 1.7	46.824	+ 19.4	51.138	+ 29.0
July	38.113	− 3.0	46.292	+ 16.0	54.971	+ 36.3
August	42.696	− .7	47.372	+ 9.5	55.114	+ 27.6
September	42.098	− 2.3	45.108	+ 4.0	51.183	+ 18.0
October	44.807	− 2.7	48.337	+ 4.8	52.300	+ 13.7
November	44.927	+ 6.5	47.690	+ 12.1	43.895	+ 2.3
December	44.098	+ 5.3	44.037	+ 4.2	40.184	− 5.7
	1919		1920		1921	
January	42.193	− 4.8	49.748	+ 13.5	41.148	− 6.8
February	32.103	− 14.3	41.055	+ 7.5	31.524	− 15.2
March	34.293	− 19.0	47.850	+ 14.0	31.055	− 26.3
April	32.712	− 3.8	38.764	+ 11.0	28.154	− 14.3
May	38.186	+ .4	39.841	+ 4.6	34.057	− 8.8
June	37.685	− 3.8	46.095	+ 16.8	34.635	− 10.4
July	43.425	+ 8.1	45.988	+ 14.6	31.047	− 21.1
August	43.613	− .4	49.974	+ 15.3	35.291	− 19.6
September	48.209	+ 10.7	50.241	+ 16.1	35.893	− 18.1
October	57.200	+ 25.5	53.278	+ 16.5	44.686	− 3.4
November	19.006	− 58.0	55.276	+ 30.5	36.805	− 19.6
December	37.235	− 12.8	53.257	+ 26.6	31.627	− 25.3
	1922		1923			
January	37.600	− 14.5	50.123	+ 18.1		
February	40.951	+ 8.8	42.160	+ 13.4		
March	50.193	+ 21.7	46.807	+ 15.2		
April	15.780	− 44.1	42.564	+ 23.7		
May	20.501	− 41.4	46.076	+ 23.7		
June	22.309	− 39.8	45.490	+ 19.3		
July	17.003	− 54.9	45.644	+ 17.9		
August	22.328	− 50.7	48.864	+ 17.2		
September	40.964	− 4.3	46.216	+ 10.7		
October	45.173	− .7	49.171	+ 11.3		
November	45.262	+ 8.7	42.946	+ 4.7		
December	46.450	+ 12.8	39.707	− 2.3		

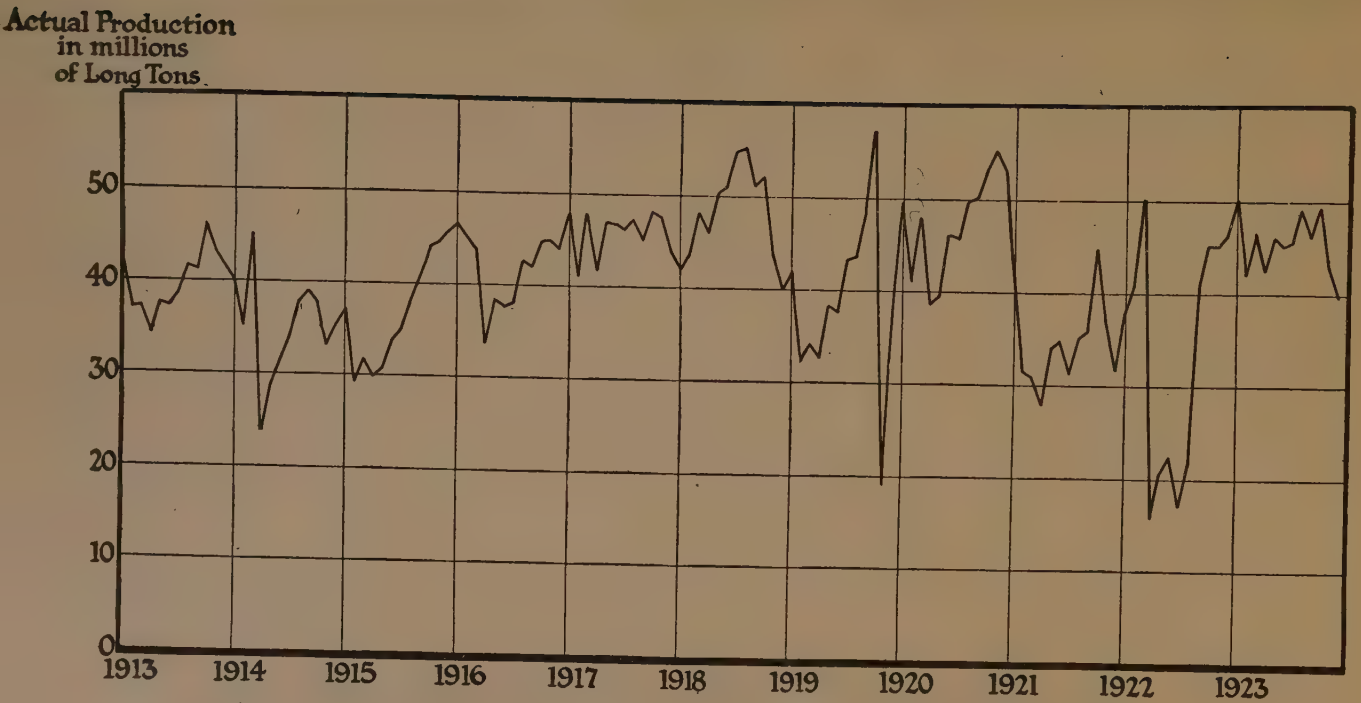


FIG. 64. — Production of Bituminous Coal in the United States, 1913–1923 (by months)

The actual monthly production figures are plotted in Fig. 64 and the cycles (corrected deviations from normal) are plotted in Fig. 65. These cycles reflect with a fair degree of accuracy the cycles in general business, though in the case of bituminous coal accidental and irregular fluctuations, which are not eliminated by the process described above, are more important than in most economic series. Thus the low production in November, 1919, and during part of 1922 was due primarily to labor troubles in the bituminous fields. Account must be taken of such irregularities in analyzing any economic series.

If cyclical changes in this series are to be compared with similar changes in other series, it is desirable to reduce the figures to a form permitting such comparison. The percentage deviations might be much more violent in one series than in another, and without a common denominator comparison would be difficult. This common denominator is afforded by the standard deviation. The monthly or annual deviations may be expressed in terms of the standard deviation as the unit of measurement, if such comparison is to be made.

The process of analysis has now been completed. We have, for the given series, the equation to the line of secular trend, and from this the normal or trend value at any given date may be computed. The seasonal variations have been measured, and indices of these variations computed. Finally, the cyclical fluctuations (plus the unmeasurable random and accidental changes) have been isolated. Further analysis of these cyclical fluctuations, looking toward the elimination of the random changes and the breaking up of the cyclical variations into simpler forms, is mathematically possible, but the greatest care is needed in applying such refined methods to economic data. A discussion of the technique of this procedure is beyond the scope of the present work.

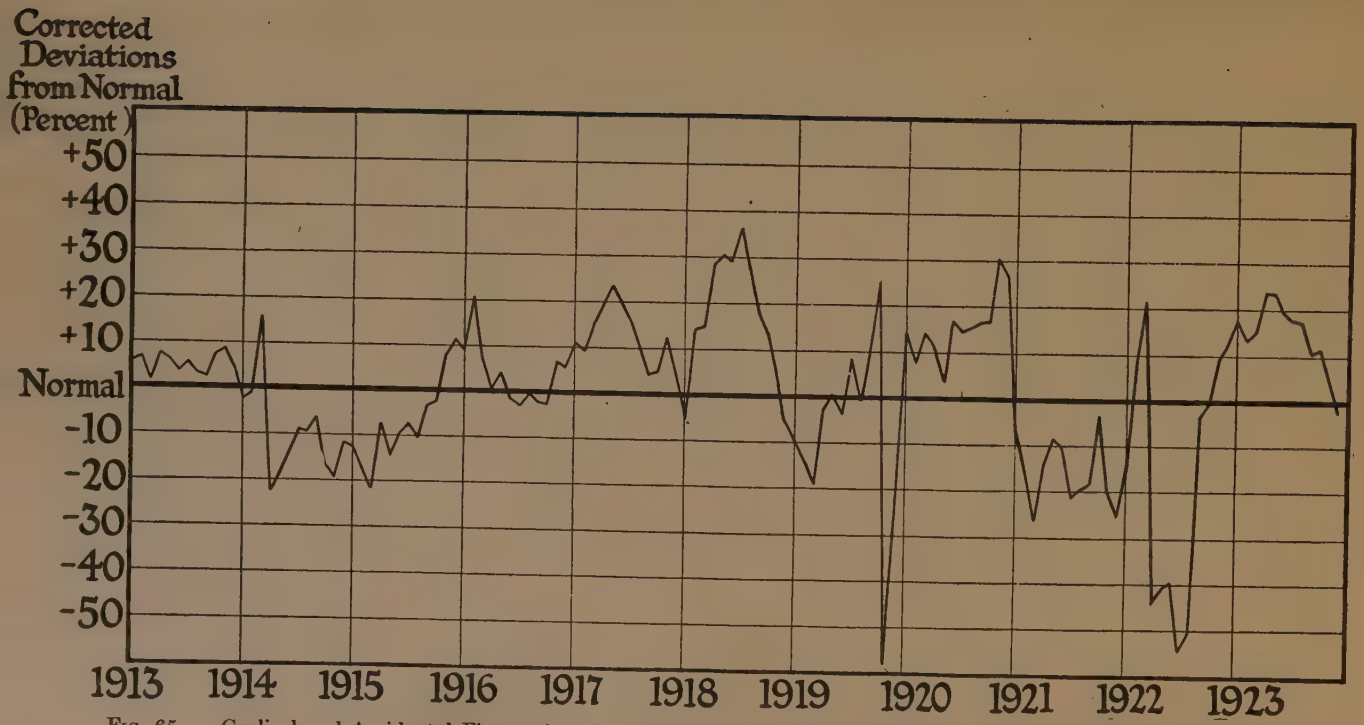


FIG. 65. — Cyclical and Accidental Fluctuations in Bituminous Coal Production in the United States, 1913–1923

General Considerations Relating to the Analysis of Series in Time

The following are certain general points which should be observed in the analysis of time series. Some of these have been noted above, and are here summarized.

1. It is a fundamental requirement that the data be homogeneous for the period covered. This applies not only to external phases, to sources, methods of quotation, etc., but to underlying economic and social conditions. Concretely, this bears directly upon the study of two of the factors described above:
   a. The representation of secular trend by a single curve is justified only if conditions affecting the trend have remained unchanged during the entire period covered.
   b. Indices of seasonal variation are to be computed only for periods marked by a fairly uniform seasonal fluctuation. This means that abnormal periods should not be included and that new indices should be constructed when underlying conditions change. The construction of such indices should be preceded by a careful preliminary study of the data and of the attendant factors in order to be sure that the conditions responsible for seasonal variation have undergone no change during the period covered.
2. In deriving the equation to a line of trend, a period containing a whole number of cycles should be included. The years selected as terminal should be approximately normal with respect to the particular series being studied.
3. For the determination of a line of trend and the calculation of indices of seasonal variation, data extending over as long a period as possible should be employed. Ten years may be suggested as the minimum period, though a much longer term of years is desirable.
4. Time series expressed in terms of dollars yet representing physical quantities should, in general, be corrected for variations in the purchasing power of the dollar before further analysis is attempted. Bank clearings, value of exports,

and value of building permits issued are examples of such series. Care should be taken that the price index employed in making the correction is an appropriate one.

REFERENCES

CHADDOCK, R. E. *Principles and Methods of Statistics* (Chap. XIII).

CRUM, W. L. *The Use of the Median in Determining Seasonal Variation.* Journal of the American Statistical Association, March, 1923.

FALKNER, HELEN D. *The Measurement of Seasonal Variation.* Journal of the American Statistical Association, June, 1924.

HALL, LINCOLN W. *Seasonal Variation as a Relative of Secular Trend.* Journal of the American Statistical Association, June, 1924.

HART, W. L. *The Method of Monthly Means for Determination of a Seasonal Variation.* Journal of the American Statistical Association, Sept., 1922.

KING, W. I. *An Improved Method for Measuring the Seasonal Factor.* Journal of the American Statistical Association, Sept., 1924. (In this article Dr. King proposes a method by which year-to-year changes in the seasonal factor may be measured.)

MOORE, H. L. *Economic Cycles: Their Law and Course.*

MOORE, H. L. *Generating Economic Cycles.*

PERSONS, W. M. *Correlation of Time Series.* (In *Handbook of Mathematical Statistics*, edited by H. L. Rietz, 151–158.)

PERSONS, W. M. *Indices of Business Conditions.* Review of Economic Statistics. Prel. Vol. I, 1919.

CHAPTER IX

INDEX NUMBERS OF PHYSICAL VOLUME

Economists and business men who are awake to the larger implications of their activities have long lamented the dearth of production statistics in the United States. For a number of staple raw materials fairly accurate production figures are available, but for the great mass of manufactured commodities no reliable data are to be had. Recent activity by governmental departments and by private agencies is tending to correct this condition, and in the future accurate and comprehensive data on the course of production will undoubtedly be available. Such data are vitally necessary to the student of prices and to the student of business cycles, as well as to business men with particular interests and economists with a more general interest in the business system.

The necessity of combining production figures in order to secure some sort of an average arises as soon as one's interest ceases to be specific, confined to one or two individual commodities. He who would study the course of general production faces much the same problem as does the student of general price movements. If the general trend of production is to be determined, or if the cyclical or seasonal swings of production are to be studied, the mass of individual figures must be reduced to the form of a single index, the significance of which may be easily comprehended. In 1919 such an index was compiled by the Price Section of the War Industries Board, under the direction of W. C. Mitchell. This index related to the production of 90 raw materials, and covered the period 1913–1918. More recently a number of similar index numbers, broader in scope and covering

longer periods, have been constructed. As illustrations of method certain of these types may be briefly described.

Production Index of the War Industries Board

The compilation of a production index by the Price Section of the War Industries Board was incidental to its major work in the field of prices. The method employed was similar to that followed in the construction of the price index. In the latter case the monthly and yearly prices of given commodities were multiplied by the amounts produced plus the amounts imported in 1917, the aggregate values secured by totaling these results being turned into price index numbers. The physical quantities were constant while prices varied. In compiling the production index the amounts of given commodities produced plus the amounts imported, year by year, were multiplied by 1917 prices. Value aggregates were secured by years, and these were turned into relatives, which constitute production index numbers.[1] Prices remaining constant throughout, the aggregate yearly values would be affected solely by changes in the amounts produced and imported.

Three sets of aggregates were worked up for these 90 raw materials, one representing yearly production times 1917 prices, one representing yearly prices times 1917 production, and the third yearly prices times yearly production. Upon the first of these the index of production was based, upon the second the index of prices, and upon the third an index of actual value changes, year by year, which represents the combined effect of changes in production and prices. These aggregates and index numbers are given in the accompanying table.

[1] The individual commodities were again weighted, the weights being proportionate to the number of manufacturing operations employed in the fabricating process. Thus raw mineral products were given more weight than raw farm products, because many of the latter undergo no manufacturing process. The weights employed were not explained in detail.

TABLE 88

Index Numbers of Production, Prices and Annual Values of 90 Raw Materials

1913–1918 [1]

Year	*Aggregate Values (in millions of dollars)*			*Index Numbers*		
	Yearly production times 1917 prices	*Yearly prices times 1917 production*	*Yearly prices times yearly production*	*Production*	*Prices*	*Prices times Prod.*
1913	$30,375	$19,973	$17,390	100	100	100
1914	30,207	19,224	16,694	99	96	96
1915	32,482	19,699	18,455	107	99	106
1916	33,700	23,363	22,785	111	117	131
1917	34,748	34,748	34,748	114	174	200
1918	35,169	38,251	39,153	116	192	225

The first index shows that the greatest increase in production came between 1914 and 1916, with much smaller increases in 1917 and 1918. The price increase is similar to that which has been revealed by other index numbers. The third column, in which both the production and price factors are variable, presents figures similar to those often quoted to show how production increased during the war. The increase in values was largely due to price changes, reflecting to only a small degree an actual increase in the physical volume of production.

The method employed above may be described by the following formula. For the index of changes in quantities produced in 1914, as compared with 1913 as base, we have

$$\frac{\Sigma q_{1914} \cdot p_{1917}}{\Sigma q_{1913} \cdot p_{1917}}.$$

This, it will be noted, is one of the aggregative types tested by Fisher in developing his "ideal" index except that the p's and the q's are interchanged. This ideal index, and the

[1] From *History of Prices During the War, Price Bulletin No. 1* of the War Industries Board, 45. The original table gives separate index numbers for five raw material groups: vegetable farm products, animal farm products, forest products, mine products, and fishery products.

simplified forms recommended for ordinary use, are as appropriate for the measurement of quantity changes as for price changes. The ideal index, when used for this purpose, would take the form

$$\sqrt{\frac{\Sigma(q_1p_0)}{\Sigma(q_0p_0)} \times \frac{\Sigma(q_1p_1)}{\Sigma(q_0p_1)}}$$

where q_0 and p_0 represent the quantities and prices of the individual commodities in the base year, while q_1 and p_1 represent quantities and prices in the given year. The procedure in the computation of such an index is identical with that employed in computing the ideal price index with prices and quantities reversed.

Day's Index Numbers of Production

Edmund E. Day has compiled two of the most useful index numbers of the physical volume of production which are currently available. The two are constructed on different principles and are designed to serve somewhat different purposes. The characteristics of the one which is termed an "unadjusted index" of the volume of production of basic materials in the United States may be first considered.

This index covers the period 1899 to date. Several changes have been made in the constituent series since the index was first constructed. The present description relates to those utilized for the period 1909–1922.

In the compilation of this index available production data were divided into four groups, those relating to agricultural products, to products of animal husbandry, to forestry products and to mineral products. A separate index number has been computed for each of these groups, as well as for the entire division of basic materials. The classification is based upon the fact that there are significant differences in the conditions affecting production in these four groups.

The methods employed in the construction of the unadjusted index are practically the same as those already explained in dealing with price index numbers. The individual production figures are first reduced to relatives, the figure for the year 1919 serving as base. (The average for the years 1919–1920 is taken as base in the case of agricultural products.) The index for each of the four minor groups is a weighted geometric average of these relatives (except in the case of forestry products, the index for which is based upon a single series, lumber cut in the United States). The weights are based upon the relative importance of the different commodities in each group, as measured by aggregate values in the base year.

The four group index numbers thus computed are based upon different numbers of commodities. For the period 1909–1922 these numbers were as follows:

Agricultural Products	16
Products of Animal Husbandry	8
Forestry Products	1
Mineral Products	13

(Figures for 16 mineral products were utilized for the later years of this period.)

The final step is the combination of these group index numbers into a single index of the production of basic materials. This final index is a weighted geometric mean of the four group index numbers, weights being proportionate to the aggregate values of production in the four fields during the base year.

By precisely similar methods an index of the volume of manufacture for the period 1914–1922 has been constructed by Day. Certain additional problems were faced in this field, particularly in the matter of weighting. For the period 1914–1918 each annual index number is a geometric mean of two indices having, respectively, 1914 and 1919 weights. For the rest of the period 1919 weights are used. This index is based upon 31 series representing commodi-

ties produced or materials consumed in current manufacture.

The values of the various unadjusted indices are given in the following table:

TABLE 89

Unadjusted Index Numbers of Production in the United States[1]

1909–1922

(1919 = 100)

Year	*Agriculture (crops)*	*Forestry*	*Animal Husbandry*	*Mining*	*Agriculture, Forestry, Animal Husbandry and Mining, Combined*	*Manufacture*
1909	86.0	128.8	82.8	74.3	83.4	...
1910	88.5	115.8	75.0	81.7	82.6	...
1911	83.8	107.1	87.0	81.0	84.8	...
1912	99.6	113.3	84.3	86.7	91.7	...
1913	88.4	111.1	83.9	93.0	87.8	...
1914	99.5	108.1	80.3	87.6	90.1	75
1915	104.9	90.4	86.1	93.3	95.3	86
1916	91.6	100.7	93.6	104.8	94.6	102
1917	97.5	96.0	93.2	112.2	98.1	105
1918	98.1	84.9	103.2	112.7	102.0	102
1919	100.0	100.0	100.0	100.0	100.0	100
1920	109.2	86.4	92.5	114.2	103.0	104
1921	90.5	78.1	94.9	91.6	92.0	80
1922	100.7	100.7	102.1	97.5	100.7	109

A comparison of these indices yields interesting results. There are pronounced and characteristic differences in the trends of the various series and there are equally marked differences in the fluctuations in production from year to year. As was pointed out in the discussion of index numbers of prices, the weather is generally the determining factor in agricultural production, while in other fields production is controlled with more immediate reference to existing prices and business conditions.

[1] These figures are from the *Review of Economic Statistics*, July, 1922, and July, 1923.

The Adjusted Index Numbers of Production

We have seen, in the analysis of time series, that the cyclical fluctuations in such series are often the objects of primary interest, particularly to one who is following changes in business conditions. This is particularly true in the study of physical volume, for changes in the volume of production and trade are features of fundamental importance in the business cycle. Methods have been explained, in the preceding chapters, by means of which we may measure such cyclical fluctuations in individual series. An obvious next step, in the study of general business conditions, is the combination of cyclical movements in a number of series into a single index. The utility of such an index of changes in the physical volume of production occurring in the course of the business cycle is evident.

When annual data are employed the construction of an index of these cyclical changes is simple. No problem of seasonal variation enters, and secular trend alone has to be taken account of. Two different methods by which this may be done present themselves. Day, in computing his "adjusted" index of the physical volume of production, has tested both methods.

The first involves the fitting of an appropriate line of trend to each of the constituent series. The actual items are then expressed as percentages of the corresponding trend values. When this has been done for each series, the final adjusted index for a given year is obtained by taking a weighted arithmetic average of these percentages for that year. Each commodity is given the same weight, in this averaging process, as in the calculation of the unadjusted index. The resulting adjusted index is in terms of relatives, but these relatives refer to a hypothetical "normal," instead of to any fixed base. This is the desired index of cyclical changes in the physical volume of production.

The alternative method is simpler. Each unadjusted

index possesses a trend which is "a composite of the persistent tendencies of the several original series upon which the unadjusted index is based." Why not measure this trend, instead of the separate original trends, and secure the adjusted index directly from the unadjusted?

The validity of the latter method may be tested by comparing the results secured by the two methods. Day has done this and has found that the results are practically identical. Accordingly, the simpler process has been employed in securing the required index numbers.

The adjusted indices for agriculture and mining, corresponding to the unadjusted values given above, are presented in the following table:

TABLE 90

Adjusted Index Numbers of Production in the United States[1]

1909–1922

Year	*Agriculture* (crops)	*Mining*
1909	98	97
1910	99	102
1911	93	98
1912	109	101
1913	95	104
1914	106	95
1915	110	98
1916	95	107
1917	100	111
1918	99	108
1919	99	93
1920	107	104
1921	88	81
1922	96	84

INDEX NUMBERS BASED ON MONTHLY VALUES

The preceding examples have dealt only with annual index numbers. By a simple elaboration of the same methods index numbers may be constructed from monthly values of series relating to the physical volume of produc-

[1] These index numbers are from the *Review of Economic Statistics*, July, 1922, and July, 1923.

tion and trade. One of the most comprehensive of these indices is that compiled and published monthly by the Federal Reserve Board.

Federal Reserve Board Index of Production in Basic Industries

This index has been constructed for the period from January, 1913, to date, being published monthly in the *Federal Reserve Bulletin*. It is based upon 22 statistical series for which monthly data are currently available. Twenty of these series are listed in Table 91 below. In compiling the final index weights are given to the various series in accordance with their relative importance, as determined from the 1919 Census of Manufactures. Data on value added in manufacture and number of men employed are used in arriving at the weights.

In the construction of this index no attempt is made to eliminate the effect of secular trend. The final index appears as a relative, with average monthly production in 1919 as base, but this base is not considered to be "normal" in the sense in which ordinates to a line of trend are taken as "normal."

Before an average is secured the various series are reduced to relatives, on the 1919 base. Seasonal variations are eliminated so far as possible, in the calculation of these relatives, by correcting the 1919 monthly averages by indices of seasonal variation. The method by which this is accomplished is worthy of note.[1]

The index numbers of seasonal variation are medians of monthly deviations from a twelve-month moving average, computed by a method which was explained in the preceding chapter. As considerable interest attaches to seasonal fluctuations in the series included, the indices secured are summarized in the following table:

[1] This method, in its particular application to this problem, is due to Frederick R. Macaulay, of the National Bureau of Economic Research.

TABLE 91

Index Numbers of Seasonal Variation in Certain Statistical Series [1]

Series	*January*	*February*	*March*	*April*	*May*	*June*	*July*	*August*	*September*	*October*	*November*	*December*
Pig iron production	99	88	103	100	102	98	100	104	101	105	102	98
Production of steel ingots	99	89	102	100	102	99	100	104	101	107	102	95
Cotton consumption	104	97	107	100	105	102	99	100	95	99	95	97
Wheat flour production	112	89	89	82	84	79	77	106	113	124	124	121
Sugar meltings	69	93	128	125	121	128	127	126	99	71	60	53
Hogs slaughtered	142	119	99	86	99	106	84	65	62	83	109	146
Cattle slaughtered	102	81	88	84	86	92	96	99	111	128	122	111
Calves slaughtered	85	73	102	121	128	114	109	96	99	100	92	81
Sheep slaughtered	104	89	91	80	83	90	100	110	120	117	110	106
Lumber cut	80	84	97	101	119	113	107	113	110	109	92	75
Bituminous coal production	107	92	102	83	92	95	97	106	106	113	104	103
Anthracite coal production	97	84	100	96	105	105	100	102	99	110	102	100
Copper production	100	93	105	103	104	102	97	100	97	103	97	99
Sole leather production	102	90	100	103	108	109	98	101	97	103	92	97
Newsprint production	106	92	101	101	101	102	98	101	95	103	98	102
Cement production	55	64	83	104	120	117	109	119	117	121	105	86
Petroleum production	98	90	102	100	102	101	105	103	99	103	98	99
Production of cigars	89	88	100	96	100	105	103	103	103	113	106	94
Production of cigarettes	95	89	99	91	99	106	114	108	104	110	99	86
Production of manufactured tobacco	94	94	107	100	102	102	100	107	106	108	95	85

[1] From the *Federal Reserve Bulletin*, Dec., 1922, 1416.

These index numbers, which are based upon data for the period 1913–19, inclusive, are quite significant in the light they throw upon seasonal fluctuations in different industrial groups.

The correction for seasonal variation is made in the figures used as bases for the various monthly relatives. The method may be illustrated by employing the figures for cotton consumption, as given in the *Federal Reserve Bulletin*. In the table which follows the actual figures for 1919, the indices of seasonal variation and the monthly averages for 1919, adjusted for seasonal variation, are presented.

TABLE 92

Application of Corrections for Seasonal Variation

(Cotton Consumption)

(1) *Month*	(2) *Cotton Consumption, 1919 (in hundreds of bales)*	(3) *Indices of seasonal variation*	(4) *Corrected monthly averages, 1919*
January	5,569	104	5,130
February	4,333	97	4,785
March	4,335	107	5,278
April	4,759	100	4,933
May	4,879	105	5,179
June	4,743	102	5,032
July	5,103	99	4,884
August	4,973	100	4,933
September	4,911	95	4,686
October	5,560	99	4,884
November	4,913	95	4,686
December	5,117	97	4,785
	59,195		59,195
Monthly average	4,933		

The monthly average for 1919 is 4,933 (hundreds of bales). If there were no seasonal variation it would be correct to assume that this, apart from trend, and accidental fluctuations, would have been the actual production

in each month of that year. But in constructing the indices of seasonal movement it has been found that January is, on the average, 4 per cent above the average for the year, February 3 per cent below, etc. Therefore, by multiplying 4,933 by the seasonal indices, we secure for the several months corrected monthly averages which indicate how the actual annual production in 1919 would have been distributed throughout the year if the force of seasonal variation were the only disturbing factor. These corrected monthly averages are used as the bases in the reduction of all actual figures to relatives.

For example, actual cotton consumption in January, 1921, amounted to 3,665 (hundreds of bales). To reduce this to a relative on the 1919 base, we divide this figure by 5,130, the corrected January figure for 1919, and multiply by 100. We secure as the relative for January, 1921, corrected for seasonal variation, the figure 71.4. To secure the February relative, the corrected February figure for 1919 is used as base. It is a great advantage of this method that the base is adjusted once and for all for seasonal variation, no later corrections being necessary.

The final index for a given month is secured readily from corrected relatives such as this, representing all the constituent series. Each relative is multiplied by its weight, the products are added, and the total divided by the sum of the weights to give the desired index. The published figures include the corrected relatives, as well as the final index of all commodities, a practice which adds materially to the usefulness of this work.

In addition to this general index of production in basic industries, the Division of Analysis and Research of the Federal Reserve Board compiles and publishes a set of useful index numbers of activity in specific lines. These include an index of agricultural movements, an index of mineral production and an index of manufacturing production. No correction for seasonal variation or secular trend

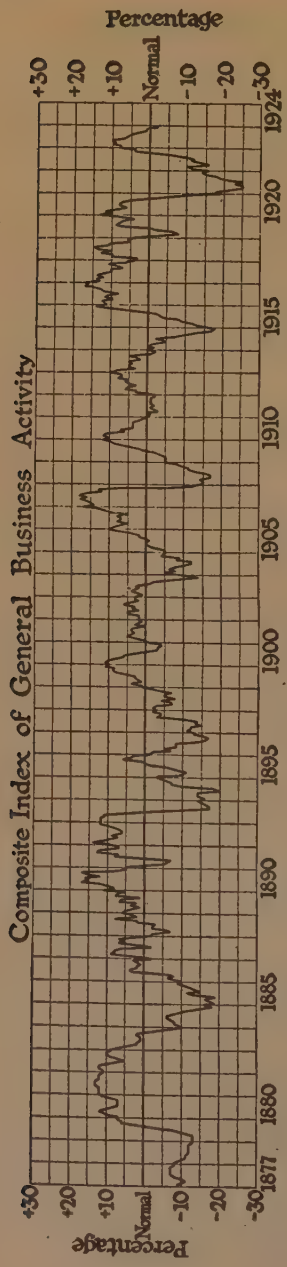


FIG. 66

(Index compiled by Statistical Division, American Telephone and Telegraph Company)

is attempted in the compilation of these index numbers. They are available, by months, for the period beginning January, 1919.

THE "GENERAL BUSINESS" INDEX OF THE AMERICAN TELEPHONE AND TELEGRAPH COMPANY

A somewhat different procedure is exemplified in the composite index of general business conditions compiled by the Statistical Division of the American Telephone and Telegraph Company. This index has been computed by months, going back to 1877, though its composition was changed a number of times during this period.

In the construction of the index as it stood at the end of 1923 eleven different series which were taken to reflect general business activity were employed. Each of these series was analyzed in the manner explained in the two preceding chapters. Lines of secular trend were fitted, the actual figures were expressed as percentages of the normal values thus secured, and

these percentages were corrected for seasonal variation. A further step had to be taken before the eleven series were combined into one. Certain series were marked by very much wider cyclical fluctuations than others, and before any averaging was attempted they all had to be reduced to comparable units. If this had not been done, the series having fluctuations of the greatest amplitude would have been over-weighted. This correction was made by expressing the deviations from normal for each series in terms of the standard deviation of that series.

The series which were combined into the final index, as of December, 1923, with their respective weights, were the following:

Series	*Weight*
Pig iron production	20
Net ton miles of freight	10
Freight car loadings	10
Steel ingot production	10
Activity of wool machinery	10
Cotton consumption	10
Paper production	10
Lumber production	5
Leather production	5
Bituminous coal production	5
Electric power production	5
Total	100

The values of this index, for the period 1877–1923 inclusive, are given in the following table. These figures are plotted in Fig. 66.

TABLE 93

Composite Index of General Business Activity, 1877–1923
Compiled by the American Telephone and Telegraph Company [1]

(The figures are in terms of percentage deviations from normal.)

	1877	1878	1879	1880	1881	1882	1883	1884	1885	1886	1887	1888
January...	−9	−8	−13	+7	+12	+11	+9	−10	−16*	−10	−2	−2
February..	−9	−9	−13	+9	+12	+11	+7	−9	−19	−8	+3	−2
March.....	−10	−10	−13	+11	+12	+12	+4	−8	−17	−6	+9	−7
April.....	−11	−11	−13	+12	+12	+12	+2	−8	−16	−8	+8	−6
May.....	−10	−11	−11	+10	+12	+10	+2	−7	−19	−7	+8	−2
June.....	−8	−12	−9	+8	+13	+8	+1	−6	−14	+3	+6	−2
July.....	−7	−12	−7	+7	+13	+5	+1	−6	−11	+4	−2	−3
August....	−7	−12	−3	+7	+13	+7	+1	−8	−15	+2	+2	+2
September.	−7	−12	+2	+7	+12	+9	+2	−10	−14	+2	+7	+2
October ...	−7	−12	+6	+7	+12	+10	+2	−12	−12	+1	+5	+8
November.	−7	−12	+6	+9	+12	+10	−2	−15	−11	+2	+7	+2
December .	−8	−13	+7	+11	+11	+9	−6	−18	−9	+3	+1	+3

	1889	1890	1891	1892	1893	1894	1895	1896	1897	1898	1899	1900
January...	+6	+11	+5	+9*	+12	−14	−3	−1	−13	−3	+1	+10
February..	+5	+8	+1	+14	+12	−16	−9	−2	−11	−2	+1	+11
March.....	+3	+10	−5	+11	+12	−14	−10	−7	−12	−3	+4	+10
April.....	+1	+12	−6	+9	+11	−14	−11	−5	−12	−7	+1	+8
May.....	+4	+17	−7	+7	+11	−14	−8	−8	−15	−7	+2	+6
June.....	+2	+15	0	+9	+5	−20	−6	−8	−13	−4	+3	+6
July.....	+7	+16	+8	+6	−2	−18	−1	−8	−14	−8	+3	+3
August....	+6	+12	+7	+6	−13	−10	0	−14	−10	−5	+6	0
September.	+1	+15	+13	+7	−17	−7	+1	−16	−3	−5	+8	−3
October ...	+8	+17	+12	+8	−17	−7	+6	−17	−4	−7	+9	−3
November.	+7	+8	+7	+10	−16	−6	+4	−16	−4	−4	+8	−4
December .	+5	+5	+7	+12	−15	−4	+3	−13	−2	−2	+11	−4

	1901	1902	1903	1904	1905	1906	1907	1908	1909	1910	1911	1912
January...	−1	+2	+3*	−10	−5	+10	+17	−14	−5*	+12	0	−2
February..	0	+2	+1	−5	−5	+9	+14	−15	−5	+10	−1	+3
March.....	+2	+2	+2	−8	−2	+7	+15	−15	−5	+12	0	+3
April.....	+5	+5	+4	−6	−1	+5	+16	−14	−5	+11	−3	+6
May.....	+4	+5	+3	−8	0	+8	+18	−17	−3	+6	−2	+5
June.....	+3	+2	+2	−10	0	+7	+16	−17	−2	+7	−1	+3
July.....	+4	+4	+3	−12	−1	+7	+18	−16	+1	+4	−2	+5
August....	+4	+3	−1	−12	0	+8	+16	−13	+2	+4	−1	+6
September.	0	+6	−2	−8	+2	+5	+12	−11	+5	+4	−1	+5
October ...	+3	+5	−6	−9	+2	+10	+13	−11	+8	+4	−1	+8
November.	+2	+2	−11	−4	+4	+11	−4	−9	+9	+2	−1	+7
December .	+1	+3	−14	−5	+7	+13	−12	−8	+10	+1	−2	+7

	1913	1914	1915	1916	1917	1918	1919	1920	1921	1922	1923	
January...	+10*	0	−16	+11	+17	+4	+4*	+13	−19*	−16	+8	
February..	+8	−2	−13	+12	+13	+3	−3	+10	−18	−11	+8	
March.....	+3	0	−12	+13	+14	+10	−8	+12	−22	−10	+10	
April.....	+5	−1	−9	+9	+12	+10	−6	+9	−25	−16*	+9	
May.....	+5	−5	−8	+12	+14	+13	−2	+6	−25	−14	+10	
June.....	+4	−4	−4	+11	+12	+10	0	+8	−23	−11	+7	
July.....	+5	−2	−3	+8	+10	+14	+7	+8	−25	−12	+6	
August....	+2	−8	−2	+12	+10	+15	+9	+7	−21	−10	+5	
September.	+3	−10	+2	+13	+7	+11	+8	+5	−19	−6	+1	
October ...	+5	−13	+5	+15	+9	+10	+7	−1	−18	−2	+1*	
November.	−1	−17	+9	+17	+11	+5	+4	−6	−15	+4	0	
December .	−2	−18	+14	+16	+6	+6	+8	−11	−14	+4	−2	

* Denotes months in which composition of curve changes.

[1] This index is published here by courtesy of the American Telephone and Telegraph Company.

An Index of the Volume of Trade

By far the most comprehensive index of physical volume is that which has been constructed by the Reports Department of the Federal Reserve Bank of New York.[1] Its scope and character may be indicated by a list of the statistical series included. The weight assigned to each series is noted.

Table 94

Statistical Series included in the Index of Trade

(Federal Reserve Bank of New York)

Productive Activity	*Weight*	
1. Consumers' Goods	8%	
2. Producers' Goods	9	
3. Factory Employment	6	
4. Motor Cars and Trucks	2	
5. Building Construction	4	
	—	29%
Primary Distribution		
6. Merchandise Car Loadings	5	
7. Other Car Loadings	2	
8. Wholesale Trade	8	
9. Exports	3	
10. Imports	2	
11. Panama Canal Traffic	1	
12. Grain Exports	1	
	—	22%
Distribution to Consumers		
13. Department Store Sales	8	
14. Chain Store Sales	3	
15. Chain Grocery Sales	6	
16. Mail Order Sales	3	
17. New Life Insurance	2	
18. Amusement Receipts	2	
19. Advertising	2	
	—	26%

[1] For a detailed description of this index see "A New Index of the Volume of Trade," Carl Snyder, *Journal of the American Statistical Association*, Dec., 1923, 49–63.

TABLE 94 — *Continued*

General Business Activity		
20. Outside Debits	8	
21. New York City Debits	5	
22. Postal Receipts	1	
23. Communication	1	
24. Electrical Power Production	2	
	—	17%
Financial Activity		
25. Shares Sold on New York Stock Exchange	2	
26. New Corporate Financing	2	
27. Grain Future Sales in Chicago	1	
28. Cotton Future Sales in New York and New Orleans	1	
	—	6%
		100%

Certain of these series are composites of more elementary data. Thus the series showing the production of consumers' goods is based upon fifteen subordinate items, and the same number are employed in deriving the series relating to producers' goods.

Certain special problems were faced in deciding upon weights for certain of these series. In general the weights were based upon value added in manufacture, value of products, or value of commodities exchanged.

In combining these series into a single index, methods already familiar to the student were employed. A "normal" was determined for each series by the fitting of an appropriate line of trend. Seasonal indices were computed in all cases in which a strong seasonal element was present. The problem of deflation was a particularly acute one, because of the many price series used. Appropriate deflating indices were readily available in some cases, but for many series deflating index numbers had to be constructed by a combination of several different series. Thus, for deflating debits to individual account outside New York, a deflating index was made up of the following constituents:

Series	*Weight*
Wholesale Commodity Prices	2
Cost of Living Index	3
Index of Average Weekly Earnings of Employees	3
Index of Stock Prices	1

Having made all these corrections in the individual series, the final combination was effected. The following values were secured for the index for the period 1919–1923:

TABLE 95

Index of the Volume of Trade, 1919–1923

(Compiled by the Federal Reserve Bank of New York)

	1919	*1920*	*1921*	*1922*	*1923*
January	97	109	90	93	109
February	97	104	92	96	110
March	95	108	91	102	113
April	101	105	92	99	109
May	106	104	91	103	110
June	108	102	92	103	107
July	110	101	91	97	100
August	107	99	95	99	102
September	107	96	94	103	100
October	109	95	92	103	106
November	105	94	92	104	105
December	105	93	93	105	100

The preceding pages have by no means given a complete account of all the current index numbers measuring volume of trade and production.[1] The primary purpose, however, has been to exemplify methods, rather than to furnish a list of available index numbers. While there is no one standard method, it will be clear that the construction of quantity index numbers requires no involved procedure. Certain knotty problems — of deflation, of determining "normal," of

[1] Reference should be made to the Index of Trade for the United States, covering the period from 1903 to date, by months, which has been compiled by Warren M. Persons and which was described in the *Review of Economic Statistics*, April, 1923; to the monthly index numbers of manufacture and mining published in the *Review of Economic Statistics;* and to the comprehensive index numbers of retail and wholesale trade published monthly in the *Federal Reserve Bulletin.*

measuring seasonal variations — are ever present, and further experimentation is needed before these can be completely solved. Existing methods are giving us, however, a series of very valuable measures of business activity which are not only materially increasing our knowledge of the economic system but are, as well, making possible more effective control of that system.

REFERENCES

DAY, EDMUND E. *An Index of the Physical Volume of Production.* Review of Economic Statistics, Sept., 1920, Jan., 1921.

DAY, EDMUND E. *The Volume of Production of Basic Materials in the United States.* Review of Economic Statistics, July, 1922.

PERSONS, W. M. *An Index of Trade for the United States.* Review of Economic Statistics, April, 1923.

SNYDER, CARL. *A New Index of the Volume of Trade.* Journal of the American Statistical Association, Dec., 1923.

SNYDER, CARL. *A New Clearings Index of Business for Fifty Years.* Journal of the American Statistical Association, Sept., 1924.

CHAPTER X

THE MEASUREMENT OF RELATIONSHIP: LINEAR CORRELATION

In discussing averages and measures of dispersion and skewness we have been dealing with methods of describing a single frequency distribution. The arrangement of the values of a single variable along a scale may be portrayed by means of these measures, which enable the central value to be determined and the character of the distribution about that central value to be described. In the analysis of time series a somewhat different problem has been faced. In such cases we are concerned with the changing values of a variable factor with the passage of time, and seek to determine the degree to which the changes in value are due to the play of different forces — the secular trend and cyclical, seasonal and accidental factors. The preceding chapters dealt with methods by which we might measure the effect upon a given series of each of these factors (with the exception of accidental fluctuations).

Certain of these methods are applicable to the problem now before us. It was found that in dealing with time series the relationship between time and the long term trend factor may be described by a definite mathematical equation. That is, trend or growth seems to be a function of time for many economic series. Where such a relationship prevails, whether it hold precisely or only approximately, there is a distinct advantage in securing a mathematical expression which describes it. A similar but much broader problem is now to be discussed. If it is possible in dealing with time series to secure a definite mathematical equation for the relation between time and the normal values of the

items in a given series, cannot the same device be employed in studying the relationship between other variables? Can we not measure, mathematically, the relation between cotton production and the price of cotton, between corn yield and rainfall, between earnings and the output of labor? If this can be done, it will place in the hands of the economist a very powerful tool, giving his methods something of the precision which attaches to the work of the physical scientist.

The Relation between Number of Taxable Personal Incomes and Motor Vehicle Registration

As a typical problem we may consider the relation between the number of taxable personal incomes and the number of motor vehicles registered, by states. The figures are given in columns (2) and (3) of Table 96.[1]

These figures are plotted in Fig. 67, each dot representing the relation between the number of taxable incomes and the number of registered motor vehicles for a given state. Such a figure is termed a "scatter diagram." It is clear from this diagram that there is a relationship between the two variables. In general, the states with a large number of taxable personal incomes are also those having a large number of motor vehicle registrations. The relationship, however, is not perfect. Two states with the same number of taxable incomes may differ quite widely in the number of registered vehicles. Thus both Kentucky and Oklahoma had 69,000 taxable personal incomes in 1921, yet the former had 126,000 motor vehicles registered, while the latter had 221,000. Were the relationship perfect a single value of the Y-variable would always be paired with a single value of the X-variable.

[1] The five leading industrial states (New York, Pennsylvania, New Jersey, Illinois and Massachusetts) have not been included. The functional relationship between personal incomes and motor vehicle ownership is not the same in industrial communities as in the country at large, and there is, accordingly, a logical justification for their exclusion.

Table 96

Taxable Personal Incomes and Motor Vehicle Registration in Forty-three States, 1921

(1) State	(2) No. of Taxable Personal Incomes, 1921 (in thousands) X	(3) No. of Motor Vehicles Registered, 1921 (in thousands) Y	(4) XY	(5) X^2	(6) Y^2
Alabama	43	82	3,526	1,849	6,724
Arizona	18	35	630	324	1,225
Arkansas	34	67	2,278	1,156	4,489
California	386	673	259,778	148,996	452,929
Colorado	70	145	10,150	4,900	21,025
Connecticut	123	137	16,851	15,129	18,769
Delaware	16	21	336	256	441
Florida	42	97	4,074	1,764	9,409
Georgia	68	131	8,908	4,624	17,161
Idaho	23	51	1,173	529	2,601
Indiana	150	400	60,000	22,500	160,000
Iowa	111	460	51,060	12,321	211,600
Kansas	89	291	25,899	7,921	84,681
Kentucky	69	126	8,694	4,761	15,876
Louisiana	68	80	5,440	4,624	6,400
Maine	44	77	3,388	1,936	5,929
Maryland	113	140	15,820	12,769	19,600
Michigan	250	477	119,250	62,500	227,529
Minnesota	125	328	41,000	15,625	107,584
Mississippi	26	65	1,690	676	4,225
Missouri	173	346	59,858	29,929	119,716
Montana	37	58	2,146	1,369	3,364
Nebraska	72	238	17,136	5,184	56,644
Nevada	10	10	100	100	100
New Hampshire	32	42	1,344	1,024	1,764
New Mexico	12	24	288	144	576
N. Carolina	44	148	6,512	1,936	21,904
N. Dakota	18	92	1,656	324	8,464
Ohio	367	720	264,240	134,689	518,400
Oklahoma	69	221	15,249	4,761	48,841
Oregon	63	118	7,434	3,969	13,924
Rhode Island	48	54	2,592	2,304	2,916
S. Carolina	25	90	2,250	625	8,100
S. Dakota	22	119	2,618	484	14,161
Tennessee	61	117	7,137	3,721	13,689
Texas	200	467	93,400	40,000	218,089
Utah	26	47	1,222	676	2,209
Vermont	18	36	648	324	1,296
Virginia	76	141	10,716	5,776	19,881
Washington	116	185	21,460	13,456	34,225
West Virginia	75	93	6,975	5,625	8,649
Wisconsin	148	341	50,468	21,904	116,281
Wyoming	22	26	572	484	676
	3,602	7,616	1,215,966	603,968	2,612,066

Our first problem is the derivation of an equation to describe this relationship which, while not perfect, is clearly existent. There is here a relationship analogous to a trend, and it is apparently a trend which can be represented by a straight line. The equation to a straight line, fitted by the method of least squares to the points on the scatter diagram, will express mathematically the *average relationship* between

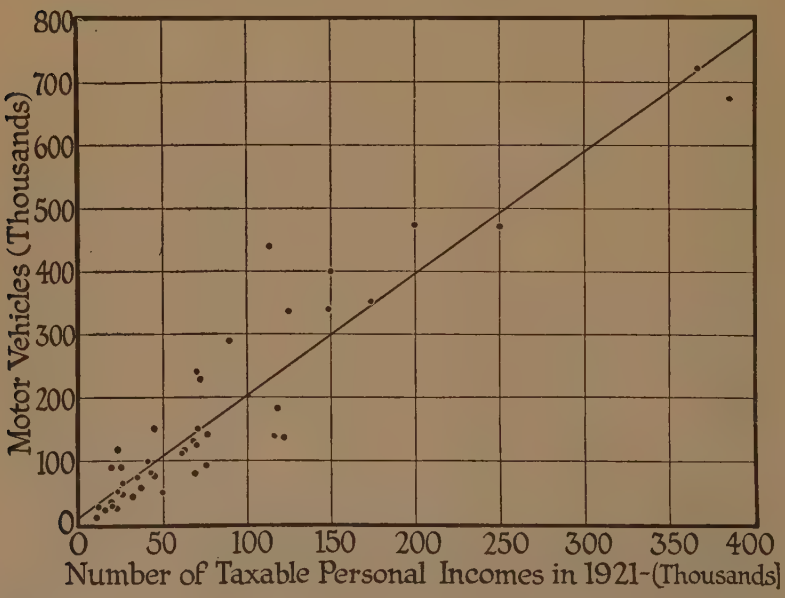


FIG. 67. — Scatter Diagram Showing the Relation between Taxable Personal Incomes and Motor Vehicle Registration, by States, in 1921, with Line of Average Relationship

these two variables. Such a line could, of course, be fitted by inspection, but a more accurate result will be obtained by the method of least squares.

This calls for the solution of the following normal equations:

$$\Sigma (Y) = na + b\Sigma(X)$$
$$\Sigma (XY) = a\Sigma(X) + b\Sigma(X^2)$$

The values required for the solution of these equations may be derived from the data as arranged in Table 96. Substituting, we have

$$7{,}616 = 43a + 3{,}602b$$

$$1{,}215{,}966 = 3{,}602a + 603{,}968b$$

Solving,

$$a = 16.92$$

$$b = 1.91$$

The required equation is

$$Y = 16.92 + 1.91X^1$$

This line is plotted in Fig. 67.

A mathematical expression has now been secured for the relation between the two variables being studied, the number of taxable personal incomes, by states, and the number of motor vehicles registered. The former is the independent or X-variable in the equation, the latter the dependent or Y-variable. This equation constitutes a measure of the functional relationship between these two variables, but it is only an expression of *average* relationship. How significant is the equation? If the relationship were perfect, and the plotted points all lay on the line describing this relationship, the equation could be used with confidence as an accurate instrument for determining the value of one variable from a value of the other. But a line with a definite equation may be fitted to points which depart very widely from it, which are widely dispersed. In such a case the equation may have the appearance of describing a precise relationship but the variation is so great that it cannot be used with confidence. It is the same problem as that which arises when an average is employed. We must know how significant the average is, how great the concentration

[1] In the chapters on correlation capital letters (W, X, Y, etc.) are used to represent original values of the variable quantities, as measured from the zero points on the scales of actual values. Small letters (w, x, y, etc.) are used to represent values of the variables expressed as deviations from their respective arithmetic means.

about it, before we may use it intelligently. So the equation of relationship between variables means little unless we know to what extent it holds in practical experience. We must have a measure of the dispersion about the line we have fitted.

In describing the frequency distribution, it has been found, the standard deviation is the best general measure of variation. It is, obviously, the measure we need in determining the reliability of the equation of average relationship. The standard deviation about this line will not only serve as a general index of the significance of this equation but will enable us to measure the degree of accuracy of estimates based upon the equation.

The Computation of the Standard Error

The standard deviation about a line of average relationship, being a measure of the accuracy of estimates, may be termed the *standard error of estimate*, or, more briefly, the *standard error*. The term *standard deviation* is generally confined to the root-mean-square deviation about the arithmetic mean. The standard error is represented by the symbol S.

In the computation of S we must know the normal value of Y which corresponds to each given value of X. By substituting the given values of X in the equation

$$Y = 16.92 + 1.191\,X$$

normal Y values may be computed. The deviations of the actual Y values from the normal may then be determined. The root-mean-square of these deviations is the required measure. A method of computation is illustrated in the following table:

TABLE 97

(1) State	(2) Motor Vehicles Registered, 1921 (*in thousands*) *Y-actual*	(3) *Y-computed*	(4) *d* (2) − (3)	(5) d^2
Alabama	82	99	− 17	289
Arizona	35	51	− 16	256
Arkansas	67	82	− 15	225
California	673	755	− 82	6,724
Colorado	145	151	− 6	36
Connecticut	137	252	− 115	13,225
Delaware	21	48	− 27	729
Florida	97	97	0	.....
Georgia	131	147	− 16	256
Idaho	51	61	− 10	100
Indiana	400	304	+ 96	9,216
Iowa	460	229	+ 231	53,361
Kansas	291	187	+ 104	10,816
Kentucky	126	149	− 23	529
Louisiana	80	147	− 67	4,489
Maine	77	101	− 24	576
Maryland	140	233	− 93	8,649
Michigan	477	495	− 18	324
Minnesota	328	256	+ 72	5,184
Mississippi	65	67	− 2	4
Missouri	346	348	− 2	4
Montana	58	87	− 29	841
Nebraska	238	155	+ 83	6,889
Nevada	10	36	− 26	676
New Hampshire	42	78	− 36	1,296
New Mexico	24	40	− 16	256
N. Carolina	148	101	+ 47	2,209
N. Dakota	92	51	+ 41	1,681
Ohio	720	719	+ 1	1
Oklahoma	221	149	+ 72	5,184
Oregon	118	137	− 19	361
Rhode Island	54	109	− 55	3,025
S. Carolina	90	65	+ 25	625
S. Dakota	119	59	+ 60	3,600
Tennessee	117	134	− 17	289
Texas	467	399	+ 68	4,624
Utah	47	67	− 20	400
Vermont	36	51	− 15	225
Virginia	141	162	− 21	441
Washington	185	239	− 54	2,916
West Virginia	93	160	− 67	4,489
Wisconsin	341	300	+ 41	1,681
Wyoming	26	59	− 33	1,089
				157,790

$$S_y = \sqrt{\frac{157{,}790}{43}}$$

$$= 60.6$$

From the calculations we secure 60.6 as the value of S_y (The symbol S_y is used as this is the standard error of the Y-variable.) This is to be interpreted in precisely the same way as the standard deviation about an arithmetic mean. Given an approximately normal distribution of items about the line of relationship, 68 per cent of all the cases will lie within a range of $\pm S$ (in this case 60.6), 95 per cent will fall within $\pm 2S$ (in this case 121.2) and 99.7 will fall within $\pm 3S$ (in this case 181.8). If there were no scatter about the line fitted to the points representing the corresponding values of X and Y, S would have a value of zero, and the value of Y could be estimated from the value of X with perfect accuracy. The less the dispersion about the line, the smaller the value of S. The value of S serves, therefore, as an indicator of the significance and usefulness of the line which describes the relationship between the two variables. The standard error, it should be noted, is expressed in the same units as the original Y-values.

The Making of Estimates

We may, for a moment, consider the significance of these results. Let us assume that, not knowing the number of motor vehicles registered in a given state, we are under the necessity of estimating it. Two methods are open to us. We may, in the first place, base the estimate upon our knowledge of the Y-variable alone. The total number of motor vehicles in the forty-three states included in the study is 7,616,000. Dividing this by 43 we have 177,116 as the average. With no specific information as to the registration in a given state, the arithmetic mean of all the state figures would be taken as the most probable value for the state in question. (The most probable value of a

series of observations is the mean of the series.) How may we judge of the accuracy of this estimate? The standard deviation of the original distribution is a measure of the degree of variation about the mean and, therefore, a measure of the accuracy of an estimate based upon the mean. If the distribution approximates the normal type, the chances are 68 out of 100 that the true value for the state in question will not differ from the mean by more than the standard deviation. The standard deviation of motor vehicle registration by states, as recorded in Table 96, is 171.4. The mean affords, therefore, a basis for a reasonable estimate, and the standard deviation affords some indication of the probabilities involved in making this estimate. (We are ignoring, of course, the possibility of basing estimates upon population, or similar factors.)

Another method of estimating the motor vehicle registration in a given state is open to us if we know the number of taxable personal incomes in that state. We know, as a result of the study described in the preceding pages, that the average relationship between motor vehicle registration and number of taxable personal incomes is described by the equation

$$Y = 16.92 + 1.91X.$$

(The unit is 1,000 for each variable, it will be recalled.)

If a state has 200,000 taxable personal incomes, it may be estimated from this equation that there are 399,000 motor vehicles registered in that state. This is the most probable value of Y as determined from the equation of average relationship. Is this estimate any better than the previous one, which took the mean Y as the most probable value? Does our knowledge of the average relationship between X and Y aid us in estimating the value of Y from a known value of X?

The answers to these questions are given by the *standard error* of Y, and by the relationship between the standard

error of Y and the standard deviation of Y. The standard error of Y (that is, the standard deviation about the line of average relationship) is 60.6. The standard deviation of Y is 171.4. Clearly the estimate made from the equation is more accurate than the estimate based upon the value of the mean Y. In the former case the odds are 68 out of 100 that the error will not exceed 60.6 or, in terms of the original units, 60,600 vehicles. When the estimate is made from the mean, the odds are 68 out of 100 that the error will not exceed 171,400 vehicles.[1] From our knowledge of the relationship between the two variables, even though that relationship is by no means constant or perfect, we are able to reduce materially the errors of estimate.

The Coefficient of Correlation

We have now secured two measures which aid us in describing the relationship between variable quantities. The first is the fundamental equation of relationship, the expression of the degree of change in one variable associated, on the average, with a given change in the other. The second is the *standard error*, the measure of the degree of "scatter" about the line of average relationship. The standard error resembles the standard deviation in that it is a measure expressed in absolute terms, in the units employed in measuring the original Y-values. This measure enables us to determine in a given case the probability that an estimate based upon the equation of relationship will fall within certain limits.

In measuring variation it has been found that an abstract

[1] In the present case, with a limited number of items and distributions which depart quite widely from the normal type, the precise probabilities cannot be so accurately determined from the values of S_y and σ_y. In particular, the presence of six or eight very large items, the rest being small in comparison, affects these two measures materially. With this qualification in the matter of interpretation we may use S_y and σ_y as useful measures of dispersion.

measure of variability is needed, one which is divorced from the absolute terms of the given problem. Such a measure is particularly needed, it was noted, when different distributions are to be compared. So, for measuring the *degree of variability*, a coefficient of variation is employed. There is need of a somewhat similar measure in connection with our present problem. We need a measure of the *degree of relationship* between two variables, an abstract coefficient which is divorced from the particular units employed in a given case. Karl Pearson has developed such a coefficient.

This measure may be explained in terms of the preceding discussion. It was found that the usefulness of estimates based upon the equation of relationship could be determined by comparing the standard error of Y (the measure of scatter about the line of relationship) with the standard deviation of Y. If the standard error be as great as the standard deviation the equation of relationship is of no use to us, but if the standard error be less than the standard deviation the accuracy of estimates may be improved by using this equation. The significance of the equation is thus indicated by the relation between the standard error and the standard deviation. But these are both in absolute terms, so that by dividing one by the other an abstract measure may be secured. Thus we might write

$$\text{Measure of correlation} = \frac{S_y}{\sigma_y}.$$

A somewhat more useful measure is secured by putting the ratio in this form:

$$\text{Measure of correlation} = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}}.$$

This measure, when used in connection with a linear equation, is called the *coefficient of correlation* and is represented by the symbol r.

A brief consideration of this formula will help to make clear the significance of r. If there is no dispersion about

the line of relationship, S_y will have a value of zero; the equation describes a perfect relationship between the two variables. In this case, as is clear from the formula, r must have a value of 1.

The maximum value of S_y is one which is equal to σ_y. Under these conditions, when the equation of relationship is of no aid in improving our estimates, the formula will give zero as the value of r. Such a value indicates that there is no relationship between the two variables; in other words, that the straight line of best fit is horizontal, passing through the mean of the Y's. It shows that there is no tendency for the high values of Y to be associated with high values of X or for high values of Y to be associated with low values of X. The two variables fluctuate in absolute independence. In such a case the deviation of each point from the fitted line is equal to its deviation from the mean, and the two root-mean-square deviations are equal, as stated.

Zero and unity are thus the limits to the value of r. The values found in practical work fall somewhere between these limits, approaching unity in cases where the degree of relationship is high. The greater the value of r, the greater the confidence that may be placed in the equation as an expression of a relation which is approximated in a high percentage of cases. In the example presented above, dealing with motor vehicle registration and number of taxable personal incomes, we have

$$r = \sqrt{1 - \frac{(60.6)^2}{(171.4)^2}}$$
$$= .935.$$

This value indicates a definite and close connection between these two variables for the states included in the sample.[1]

The coefficient of correlation may be made somewhat more significant by giving it the sign of the constant b in

[1] The high value of r in this case is in part due to the presence of six or eight items which are much larger than the main group. The effect of such exceptional cases upon the coefficient of correlation is discussed in more detail below.

the equation of relationship. This sign indicates whether the slope of the line is positive or negative and, when attached to r, enables us to tell whether the relationship is direct or inverse. Thus in the present case high values of one value are paired with high values of the other. The correlation is positive and the coefficient should be written + .935. When cotton production and prices are correlated the relationship is an inverse one, for high values of one variable are generally associated with low values of the other.

The measurement of relationship in a given case is completed when we have secured the three measures described. The *equation of average relationship* is an expression of the underlying law connecting the two variables, if such a law may be assumed. The *standard error* measures the variation, in absolute terms, about the line of relationship. The *coefficient of correlation* is an abstract measure of the degree to which the average relationship actually holds in practice.

Details of Calculation

In the preceding section the attempt has been made to explain the various measures necessary in studying the relationship between variable quantities without introducing a detailed explanation of procedure. We may now return to a consideration of the details of calculation, including certain methods by which this calculation may be reduced to a minimum.

The procedure followed in the preceding illustration is a logical one to employ in deriving the three required values. This method is capable of general application, but the labor involved may be materially reduced by taking advantage of a short-cut method of deriving S_y. This method may be first explained with reference to data of the type dealt with above. And, for the present, the discussion will be confined to cases in which the relationship between variables may be described by a straight line.

The first problem is the derivation of the equation of relationship. A line of the type

$$Y = a + bX$$

is fitted by the method of least squares.

The next step is the computation of S_y^2, the square of the standard error. This was done in the above illustration by measuring the deviation of each individual observation from the fitted line, and getting the mean-square of these deviations. It may be shown [1] that this value can be derived from the following equation:

[1] The standard error is computed from the formula

$$S_y^2 = \frac{\Sigma (d^2)}{N}$$

where d represents a single deviation from the fitted line, or the difference between the actual and the computed value of Y in a given case. The latter is derived from the equation

$$Y_c = a + bX.$$

(The symbol Y_c is used to represent the computed value of Y.)

If we let Y represent the actual value we have, for each residual,

$$d = Y_c - Y$$

or

$$d = a + bX - Y. \tag{1}$$

There will be as many equations of this type as there are points. Multiplying each one by d, and adding, we have

$$\Sigma(d^2) = a\Sigma(d) + b\Sigma(dX) - \Sigma(dY). \tag{2}$$

But, since the line was fitted by the method of least squares,

$$\Sigma(d) = 0$$
$$\Sigma(dX) = 0$$

(For proof of this see Appendix A.)
and, therefore,

$$\Sigma(d^2) = -\Sigma(dY) \tag{3}$$

Returning again to equation (1), we may multiply throughout by Y, and add, securing

$$\Sigma(dY) = a\Sigma(Y) + b\Sigma(XY) - \Sigma(Y^2). \tag{4}$$

Substituting the equivalent of $\Sigma(dY)$ in equation (3) we have

$$\Sigma(d^2) = \Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY) \tag{5}$$

from which the given formula for S_y^2 is derived.

$$S_y^2 = \frac{\Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY)}{N}$$

The quantities a and b are the constants in the equation to the fitted straight line. The other values relate to the original observations. Substituting in this equation the necessary values, taken from Table 96 and from page 367 above, we have[1]

$$S_y^2 = \frac{2{,}612{,}066 - (16.92 \times 7{,}616) - (1.91238 \times 1{,}215{,}966)}{43}$$

$$= \frac{157{,}814}{43}$$

$$= 3{,}670$$

$$S_y = 60.6$$

From this point the procedure may follow that already described, r being computed from the formula

$$r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}}.$$

The coefficient r may be secured, however, without computing S as an intermediate value. The above formula for r may be reduced to

$$r^2 = \frac{a\Sigma(Y) + b\Sigma(XY) - Nc_y^2}{\Sigma(Y^2) - Nc_y^2}$$

where c_y is the difference between the mean Y and the origin employed in the calculations.[2] If the origin is zero

[1] For this calculation the value of b is given to a greater number of decimal places than in the equation as first presented.

[2] The formula

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2}$$

may be written

$$r^2 = 1 - \frac{\Sigma(d^2)}{\Sigma(y^2)}$$

in which y refers to deviations from the arithmetic mean of the Y's. But

$$\frac{\Sigma(y^2)}{N} = \frac{\Sigma(Y^2)}{N} - c_y^2$$

where Y represents a deviation from an arbitrary origin (in this case zero on the

on the original Y scale, c_y will be equal to the arithmetic mean of the Y's.

In the present case, using the data of Table 96, we have

$$c_y = \frac{7{,}616}{43} = 177.116.$$

The other values are the same as those employed above in computing S_y. Substituting in the formula, we have

$$r^2 = \frac{1{,}105{,}337.98}{1{,}263{,}152.7}$$

$$= .875$$

$$r = .935$$

In effect, then, the labor of fitting a straight line by the method of least squares gives us practically all the values needed in securing S and r, the two other measures necessary for a complete description of the relationship between two variable quantities. The only additional values required are $\Sigma(Y^2)$ and c_y.

The Construction of a Correlation Table

In the example presented above we had only forty-three observations. With a larger number it becomes practically impossible to retain the individual values in the study of relationships. These individual items must be grouped in

---

original scale) and c_y represents the difference between this origin and the mean of the Y's.

Therefore

$$r^2 = 1 - \frac{\Sigma\,(d^2)}{\Sigma\,(Y^2) - Nc_y^2}$$

Substituting in this equation the equivalent of $\Sigma\,(d^2)$, as given in the footnote on page 376,

$$r^2 = 1 - \frac{\Sigma\,(Y^2) - a\Sigma\,(Y) - b\Sigma\,(XY)}{\Sigma\,(Y^2) - Nc_y^2}.$$

Simplifying,

$$r^2 = \frac{a\Sigma\,(Y) + b\Sigma\,(XY) - Nc_y^2}{\Sigma\,(Y^2) - Nc_y^2}.$$

significant classes, and all computations must be based upon these grouped data. This means, merely, that we must handle data organized in frequency distributions. Since we are dealing with two variables, however, the simple frequency table must be modified to meet the needs of the present problem. Such a modified frequency table, arranged to facilitate the computation of the values needed in studying relationship, is termed a *correlation table.*

As a typical problem, involving the construction of such a table, we may consider the relation between the discount rates of Federal Reserve banks on a standard type of commercial paper and the discount rates of commercial banks upon similar paper. Since this paper may be rediscounted at the Federal Reserve banks by the member banks, some degree of relationship between the rates may be expected. Our present object is the measurement of that relationship.

The first step is the tabulation of the original observations. Monthly values of each variable [1] were secured for each of the twelve Federal Reserve cities over a period of 42 months, from July 1, 1920, to December 1, 1923. In the process of tabulation the items must be combined so that a Federal Reserve bank discount rate is paired with the corresponding rate charged by the commercial banks of the same city. Fig. 68 illustrates the method of tabulation.

[1] The discount rates of the Federal Reserve bank relate to trade acceptances maturing within 90 days. The commercial bank rates are those charged on customers' prime commercial paper maturing in from 30 to 60 days (in some cases from 30 to 90 days). The customary rate over a given 30 day period was taken as of the middle of that period.

X-Federal Reserve Bank Discount Rate

Y-Commercial Bank Rates	3.75 to 4.24%	4.25 to 4.74%	4.75 to 5.24%	5.25 to 5.74%	5.75 to 6.24%	6.25 to 6.74%	6.75 to 7.24%	Totals
7.75 to 8.24%					𝍸 /	𝍸 /	𝍸 //	4
7.25 to 7.74%					𝍸 𝍸 //	𝍸 𝍸 ////	𝍸 /	17
6.75 to 7.24%			𝍸	////	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 ///	𝍸 ////	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 /	117
6.25 to 6.74%		//	𝍸 ////	𝍸 𝍸	𝍸 𝍸 𝍸 𝍸 𝍸 //	/	///	47
5.75 to 6.24%	/	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸

TABLE 98

Correlation Table, Discount Rates of Federal Reserve Banks and Discount Rates of Commercial Banks

X–Federal Reserve Banks' Discount Rates (per cent)

Y–Discount Rates of Commercial Banks (per cent)	*Class-Intervals*						3.75–4.24	4.25–4.74	4.75–5.24	5.25–5.74	5.75–6.24	6.25–6.74	6.75–7.24	Totals
		Mid-points					4.00	4.50	5.00	5.50	6.00	6.50	7.00	
			f				24	227	48	20	123	20	42	504
				d'			0	1	2	3	4	5	6	
					fd'		0	227	96	60	492	100	252	1227
						$f(d')^2$	0	227	192	180	1968	500	1512	4579
	7.75–8.24	8.00	4	8	32	256					1	1	2	
	7.25–7.74	7.50	17	7	119	833					7	9	1	
	6.75–7.24	7.00	117	6	702	4212			5	4	63	9	36	
	6.25–6.74	6.50	47	5	235	1175		2	9	10	22	1	3	
	5.75–6.24	6.00	156	4	624	2496	1	90	29	6	30			
	5.25–5.74	5.50	126	3	378	1134	11	110	5					
	4.75–5.24	5.00	34	2	68	136	10	24						
	4.25–4.74	4.50	3	1	3	3	2	1						
	Totals		504		2161	10245								

In this table, it will be noted, an arbitrary origin is employed for each variable, and the class-interval unit is used in the calculations. We may employ the symbols X' and Y' to represent the deviations from the arbitrary origin (which is located at point 4,4 on the original scales).

COMPUTATION OF MEASURES OF RELATIONSHIP

From this correlation table all the values needed in fitting a straight line to the data, and in computing the measures S and r, may be derived. The quantities $\Sigma(X')$, $\Sigma(X')^2$, $\Sigma(Y')$ and $\Sigma(Y')^2$ are computed by methods already familiar to the student. The product of the paired values $\Sigma(X'Y')$ may be computed directly from the correlation table, but it is perhaps simpler for the beginner to re-arrange the data

in columnar form, as in the following table. When the figures are disposed in this way one line is employed for each compartment of the original correlation table in which items have been recorded.

TABLE 99

Discount Rates of Federal Reserve Banks and Discount Rates of Commercial Banks

Computation of Values Required in Curve Fitting

(1)	(2)	(3)	(4)
X'	Y'	f	$f(X'Y')$
0	1	2	0
0	2	10	0
0	3	11	0
0	4	1	0
1	1	1	1
1	2	24	48
1	3	110	330
1	4	90	360
1	5	2	10
2	3	5	30
2	4	29	232
2	5	9	90
2	6	5	60
3	4	6	72
3	5	10	150
3	6	4	72
4	4	30	480
4	5	22	440
4	6	63	1,512
4	7	7	196
4	8	1	32
5	5	1	25
5	6	9	270
5	7	9	315
5	8	1	40
6	5	3	90
6	6	36	1,296
6	7	1	42
6	8	2	96
			6,289

The values required in fitting a straight line and in computing the standard error and the coefficient of correlation are:

$$N = 504 \qquad \Sigma(X')^2 = 4{,}579$$
$$\Sigma(X') = 1{,}227 \qquad \Sigma(X'Y') = 6{,}289$$
$$\Sigma(Y') = 2{,}161 \qquad \Sigma(Y')^2 = 10{,}245$$

The equation to the best fitting straight line is found to be

$$Y' = 2.7155 + .6458X'.$$

Substituting in the formula

$$S_y^2 = \frac{\Sigma(Y')^2 - a\Sigma(Y') - b\Sigma(X'Y')}{N}$$

we have

$$S_y^2 = \frac{10{,}245 - (2.7155 \times 2{,}161) - (.6458 \times 6{,}289)}{504}$$
$$= .6257$$
$$S_y = .791$$

To determine the value of the coefficient of correlation we have only to substitute the proper values in the equation

$$r^2 = \frac{a\Sigma(Y') + b\Sigma(X'Y') - Nc_y^2}{\Sigma(Y')^2 - Nc_y^2}.$$

When this is done we have

$$r^2 = \frac{(2.7155 \times 2{,}161) + (.6458 \times 6{,}289) - (504 \times 18.38437)}{10{,}245 - (504 \times 18.38437)}$$
$$= \frac{663.9}{979.27}$$
$$= .67795$$
$$r = +.82$$

All these calculations have been carried through in class-interval units, with reference to an origin at point 4, 4 on the original scales. The value of r is not affected by this fact, but the estimating equation and the standard error should be corrected.

The value of S_y, in class-interval units, is .791. Since the class interval of the Y-variable is .5%, we have, in original units,

$$S_y = .5\% \times .791$$
$$= .40\%.$$

The equation may be corrected in a similar fashion. The class interval being .5% for both X and Y, we may write

$$2Y' = 2.7155 + .6458(2X')$$

or

$$Y' = 1.3577 + .6458X'$$

which is the equation in terms of original units.

In changing the origin, we know that $Y' = Y - 4$ and $X' = X - 4$. Therefore

$$Y - 4 = 1.3577 + .6458(X - 4)$$

or

$$Y = 2.7745 + .6458X.$$

We have now the three values required for determining the relationship between Federal Reserve discount rates and corresponding commercial bank rates, during the period covered. The equation describes the average relationship, the standard error serves as a measure of the reliability of estimates based upon this equation, and the coefficient of correlation serves as an abstract measure of the degree of relationship between the two variables.

The significance of the standard error, S_y, is brought out graphically in Fig. 69. The line of average relationship has been drawn on this scatter diagram, and what may be called "zones of estimate" have been marked out about this line. Within the zone having a width equal to $2S$, centering at the fitted straight line, 68 per cent of all the points should fall, on the assumption that the distribution is normal. Within the zone having a width equal to $6S$, centering at the fitted straight line, 99.7 per cent of all the points should fall, on the same assumption. The smaller the value of S the narrower these zones would be, and hence the more accurate would be the estimates which are based upon the equation of average relationship.

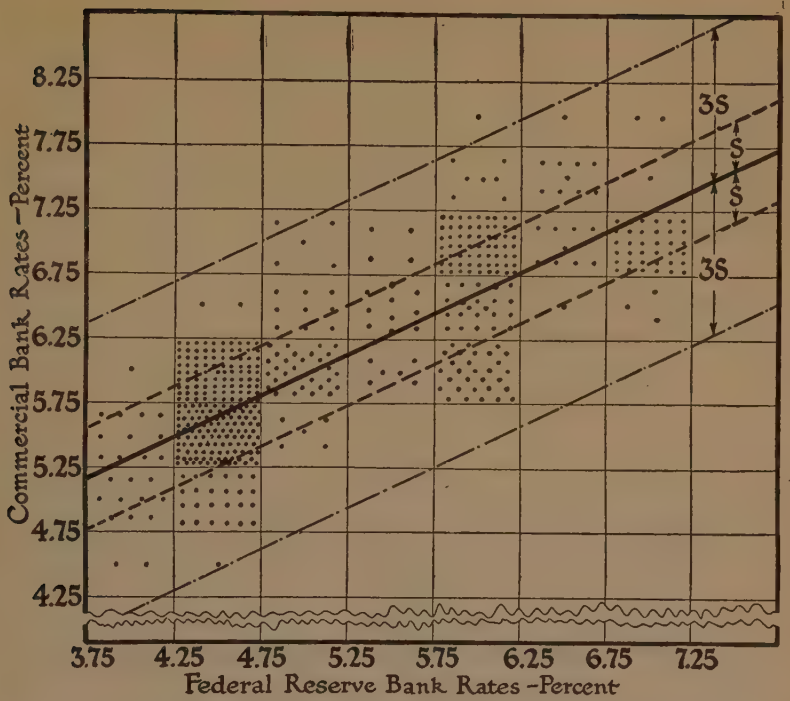


FIG. 69. — Scatter Diagram of Federal Reserve and Commercial Bank Rates, with Line of Average Relationship and Zones of Estimate

THE PRODUCT-MOMENT FORMULA FOR THE COEFFICIENT OF CORRELATION

In the preceding examples the coefficient of correlation has been computed from the formula

$$r^2 = \frac{a\Sigma(Y) + b\Sigma(XY) - Nc_y^2}{\Sigma(Y^2) - Nc_y^2}$$

which is derived directly from the method of fitting a straight line by least squares. The usual formula differs somewhat from this, and it is advisable that the student be familiar with it.

When a straight line is fitted to data, the origin being at the point of averages, the two normal equations

$$\Sigma(Y) = na + b\Sigma(X)$$

$$\Sigma(XY) = a\Sigma(X) + b\Sigma(X^2)$$

reduce to $$\Sigma(xy) = b\Sigma(x^2)$$

for $$\Sigma(x) = 0 \text{ and } \Sigma(y) = 0$$

The slope, b, is the only constant required, and this may be computed from the relationship

$$b = \frac{\Sigma(xy)}{\Sigma(x^2)}.$$

Under the same conditions the formula

$$r^2 = \frac{a\Sigma(Y) + b\Sigma(XY) - Nc_y^2}{\Sigma(Y^2) - Nc_y^2}$$

reduces to $$r^2 = \frac{b\Sigma(xy)}{\Sigma(y^2)}$$

for $c_y = 0$ when the deviations are measured from the mean of the Y's. Substituting for b its equivalent, as just determined, we have

$$r^2 = \frac{\Sigma(xy) \cdot \Sigma(xy)}{\Sigma(y^2) \cdot \Sigma(x^2)}.$$

But $\Sigma(y^2) = N\sigma_y^2$ and $\Sigma(x^2) = N\sigma_x^2$. Therefore

$$r^2 = \frac{\Sigma(xy) \cdot \Sigma(xy)}{N^2\sigma_y^2\sigma_x^2}$$

and

$$r = \frac{\Sigma(xy)}{N\sigma_x\sigma_y}$$

in which x and y refer to deviations from an origin at the point of averages.

This formula may be expressed

$$r = \frac{p}{\sigma_x\sigma_y}$$

in which

$$p = \frac{\Sigma(xy)}{N}$$

The quantity p is the *mean product* of the paired values of x and y.

The computation of the coefficient of correlation from this formula proceeds along lines somewhat different from

those outlined above. As we have seen, both the arithmetic mean and the standard deviation may be readily computed by the selection of an arbitrary origin from which all deviations are measured, a later correction being made to offset the error involved in using this arbitrary origin. Similarly, the mean product p may be computed by a short method, requiring the use of assumed means and the application of a correction at the end of the process.

If x' and y' represent deviations from points arbitrarily selected as assumed means, while p' represents the mean product of such deviations, then

$$p' = \frac{\Sigma(x'y')}{N}.$$

The computation of p' is not difficult, for deviations may be measured from central points, and may be expressed in class-interval units. Having p' we may secure the true mean product from the formula

$$p = p' - c_x c_y$$

in which c_x and c_y represent the differences between the true and assumed means of the x's and y's, respectively.[1]

[1] The following is a proof of this relationship:

x' = deviation of any point from assumed mean of x's
x = deviation of same point from true mean of x's
c_x = difference between true and assumed means of x's
y' = deviation of same point from assumed mean of y's
y = deviation of same point from true mean of y's
c_y = difference between true and assumed means of y's

$$x' = x + c_x$$
$$y' = y + c_y$$
$$x'y' = (x + c_x)(y + c_y) = xy + c_x y + c_y x + c_x c_y.$$

For the sum of all such products for N points, we have

$$\Sigma(x'y') = \Sigma(xy) + c_x \Sigma(y) + c_y \Sigma(x) + N c_x c_y.$$

But $\Sigma(y) = 0$ and $\Sigma(x) = 0$.

Therefore

$$\Sigma(x'y') = \Sigma(xy) + N c_x c_y$$
$$\frac{\Sigma(x'y')}{N} = \frac{\Sigma(xy)}{N} + c_x c_y$$
$$\frac{\Sigma(xy)}{N} = \frac{\Sigma(x'y')}{N} - c_x c_y$$

or $p = p' - c_x c_y$.

The Product-Moment Method, Ungrouped Data

This method may be illustrated with reference, first, to ungrouped data, using the figures for personal incomes (X) and motor vehicle registration (Y), by states. The values required for this computation, as given in Table 96, are

$$N = 43$$
$$\Sigma(X) = 3{,}602$$
$$\Sigma(Y) = 7{,}616$$
$$\Sigma(X^2) = 603{,}968$$
$$\Sigma(Y^2) = 2{,}612{,}066$$
$$\Sigma(XY) = 1{,}215{,}966$$

The mean product can be computed from the formula

$$p = \frac{\Sigma(xy)}{N} = \frac{\Sigma(x'y')}{N} - c_x c_y$$

We may select as arbitrary origin the actual origin on the two original scales. Hence we have

$$p = \frac{\Sigma(XY)}{N} - c_x c_y.$$

For the two standard deviations

$$\sigma_x = \sqrt{\frac{\Sigma(X^2)}{N} - c_x^2}$$

$$\sigma_y = \sqrt{\frac{\Sigma(Y^2)}{N} - c_y^2}.$$

(When the arbitrary origin is at zero on the original scales, the symbol X corresponds to x' and Y corresponds to y', as used in the formulas.)

These measures may be computed readily from the values secured from Table 96:

$$c_x = \frac{3{,}602}{43} = 83.767 \qquad c_y = \frac{7{,}616}{43} = 177.116$$

$$c_x^2 = 7{,}016.910 \qquad c_y^2 = 31{,}370.077$$

$$p = \frac{1{,}215{,}966}{43} - 14{,}836.476$$

$$= 13{,}441.803$$

$$\sigma_x = \sqrt{\frac{603{,}968}{43} - 7{,}016.910} \qquad \sigma_y = \sqrt{\frac{2{,}612{,}066}{43} - 31{,}370.077}$$

$$= 83.84 \qquad 171.41$$

Solving for the coefficient of correlation,

$$r = \frac{p}{\sigma_x \sigma_y} = \frac{13{,}441.803}{83.84 \times 171.41} = +\ .935$$

The equation to the straight line which describes the average relationship between X and Y may be derived from the values required for the preceding calculations. When the origin is at the point of averages this equation may be written

$$y = r\frac{\sigma_y}{\sigma_x}x.$$

Substituting the proper values, we have

$$y = +\ .935\,\frac{171.41}{83.84}\,x$$
$$= 1.91x.$$

This, it will be noted, is the equation secured by the method of least squares. The constant term representing the y-intercept disappears, since the origin is at the point of averages, through which the least squares line must pass.[1]

[1] That the formula $y = r\frac{\sigma_y}{\sigma_x}x$ is equivalent to the formula based upon the method of least squares may be readily demonstrated. When the line passes through the point of averages, the equation, $Y = a + bX$, becomes $y = bx$. But $b = \frac{\Sigma(xy)}{\Sigma(x^2)}$. We may write, accordingly, $y_c = \frac{\Sigma(xy)}{\Sigma(x^2)}x$

This is equivalent to

$$y_c = r\frac{\sigma_y}{\sigma_x}x$$

for the latter may be written

(1) $y_c = \frac{\Sigma(xy)}{N\sigma_y\sigma_x}\cdot\frac{\sigma_y}{\sigma_x}x$ (3) $y_c = \frac{\Sigma(xy)}{N\sqrt{\frac{\Sigma(x^2)}{N}}\cdot\sqrt{\frac{\Sigma(x^2)}{N}}}x$

(2) $y_c = \frac{\Sigma(xy)}{N\sigma_x\cdot\sigma_x}x$ (4) $y_c = \frac{\Sigma(xy)}{\Sigma(x^2)}x$

(The symbol y_c is employed for the computed value of y, in these equations, to avoid confusion with the actual y's which appear in the right-hand members of the equations.)

When the product-moment method is employed in computing the coefficient of correlation and in determining the equation of regression, the standard error, S_y, may be derived by a simple change in the formula first presented for r. From the expression

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2}$$

we may secure the formula

$$S_y = \sigma_y\sqrt{1 - r^2}$$

which enables us to compute S_y, if we have the values of σ_y and r. In the present case,

$$S_y = 171.41\sqrt{1 - (.935)^2}$$
$$= 60.6$$

The Product-Moment Method, Classified Data

The product-moment method is also applicable to cases in which it is necessary to construct a double frequency or correlation table. The procedure is shown in detail in Table 100.

This table is identical with that previously presented for the same data, except that a different arbitrary origin has been selected.

The value 5.5 is adopted as the assumed mean of the X's, and the value 6.5 as the assumed mean of the Y's. Deviations are measured in class-interval units from this origin. In each compartment of the correlation table there are three figures, involved in the computation of $\Sigma(x'y')$. The figure in the center indicates the number of items falling in that compartment. Thus there are seven pairs having X values between 5.75 and 6.25 (mid-point 6.0) and Y values between 7.25 and 7.75 (mid-point 7.5). For each of these pairs x' (the deviation from the assumed mean of the X's) is + 1, in class-interval units, and y' (the deviation from the assumed mean of the Y's) is + 2, in class-interval units. For each pair, therefore, $x'y' = + 2$.

TABLE 100

Correlation Table Showing the Relation Between Federal Reserve Bank Discount Rates and the Discount Rates of Commercial Banks, and Illustrating the Computation of the Coefficient of Correlation

Y—Commercial Banks' Discount Rates (Per cent)	Class-intervals	Mid-points	f	d′	fd′	f(d′)²	X — Federal Reserve Banks' Discount Rates (Per cent) 3.75—4.24	4.25—4.74	4.75—5.24	5.25—5.74	5.75—6.24	6.25—6.74	6.75—7.24	Totals	Σ(x′y′)
							4.0	4.5	5.0	5.5	6.0	6.5	7.0		
			f				24	227	48	20	123	20	42	504	
				d′			−3	−2	−1	0	+1	+2	+3		
					fd′		−72	−454	−48	0	+123	+40	+126	−285	
						f(d′)²	216	908	48	0	123	80	378	1753	
	7.75—8.24	8.0	4	+3	+12	36					+3 1 (+3)	+6 1 (+6)	+9 2 (+18)		+27
	7.25—7.74	7.5	17	+2	+34	68					+2 7 (+14)	+4 9 (+36)	+6 1 (+6)		+56
	6.75—7.24	7.0	117	+1	+117	117			−1 5 (−5)	0 4 (0)	+1 63 (+63)	+2 9 (+18)	+3 36 (+108)		+184
	6.25—6.74	6.5	47	0	0	0		0 2 (0)	0 9 (0)	0 10 (0)	0 22 (0)	0 1 (0)	0 3 (0)		
	5.75—6.24	6.0	156	−1	−156	156	+3 1 (+3)	+2 90 (+180)	+1 29 (+29)	0 6 (0)	−1 30 (−30)				+182
	5.25—5.74	5.5	126	−2	−252	504	+6 11 (+66)	+4 110 (+440)	+2 5 (+10)						+510
	4.75—5.24	5.0	34	−3	−102	306	+9 10 (+90)	+6 24 (+144)							+234
	4.25—4.74	4.5	3	−4	−12	48	+12 2 (+24)	+8 1 (+8)							+32
	Totals		504		−359	1285									+1225

TABLE 100, *Continued*

$A.M_x = 5.5$

$$c_x = \frac{-285}{504}$$
$$= -.565$$
$$c_x^2 = .319$$
$$s_x^2 = \frac{\Sigma f(d')^2}{N}$$
$$= \frac{1753}{504}$$
$$= 3.478$$
$$\sigma_x^2 = s^2 - c^2$$
$$= 3.478 - .319$$
$$= 3.159$$
$$\sigma_x = \sqrt{3.159}$$
$$= 1.777$$

$A.M_y = 6.5$

$$c_y = \frac{-359}{504}$$
$$= -.712$$
$$c_y^2 = .507$$
$$s_y^2 = \frac{1235}{504}$$
$$= 2.450$$
$$\sigma_y^2 = 2.450 - .507$$
$$= 1.943$$
$$\sigma_y = 1.394$$

$$p = \frac{\Sigma(x'y')}{N} - c_x c_y$$
$$= \frac{1225}{504} - .402$$
$$= 2.4305 - .402$$
$$= 2.0285$$
$$r = \frac{p}{\sigma_x \sigma_y}$$
$$r = \frac{2.0285}{1.777 \times 1.394}$$
$$= \frac{2.0285}{2.477}$$
$$r = +.82$$
$$M_x = 5.218$$
$$M_y = 6.144$$

NOTE: The class-interval unit has been employed in all the computations shown on this table.

This figure appears at the top of the compartment. But there are seven pairs in this compartment, so the sum of $x'y'$ for this group is $+14$. This figure appears in parentheses at the bottom of the compartment. To secure $\Sigma(x'y')$ it is necessary to add algebraically the values secured in this way for all compartments. The addition is first carried out for the different rows, the subtotals being given in the column at the right of the table. It is found that $\Sigma(x'y') = +1225$, in class-interval units.

The values of c_x and c_y have been secured in computing the standard deviations by methods already familiar. These values, also in class-interval units, are

$$c_x = -.565 \qquad c_y = -.712$$

We have, therefore

$$\frac{\Sigma(xy)}{N} = \frac{\Sigma(x'y')}{N} - c_x c_y$$
$$= \frac{1225}{504} - .402$$
$$= 2.0285$$

This is the value of p, the mean product, in class-interval units. Proceeding,

$$r = \frac{\Sigma(xy)}{N\sigma_x\sigma_y}$$

$$= \frac{p}{\sigma_x \sigma_y}$$

$$= \frac{2.0285}{1.777 \times 1.394}$$

$$= +.82$$

In computing r, both the numerator and denominator of the final fraction (the mean product and the two standard deviations) are in class-interval units. Since this is true, r may be computed directly without reducing the figures to the original units. The entire operation, therefore, is carried on in simple class-interval units.

In deriving the equation to the straight line which describes the average relationship between x and y from the formula

$$y = r\frac{\sigma_y}{\sigma_x}x$$

σ_y and σ_x should be expressed in units of the original scale.[1] This is done by multiplying the present values by the class-intervals.

$$\sigma_x(\text{in original units}) = 1.777 \times .50 = .8885$$
$$\sigma_y(\text{in original units}) = 1.394 \times .50 = .697$$

Substituting the given values in the formula we have

$$y = .82\frac{.697}{.8885}x$$

$$= .64x$$

The Lines of Regression

In the above discussion certain terms ordinarily employed in the treatment of correlation have been purposely omitted. Several of these should be explained.

[1] When the class-intervals happen to be the same, as in the present case, the change is not necessary, as the relation between numerator and denominator is not altered. In practice it is advisable always to express the two standard deviations in original units at this stage in the calculations.

The equation to the line of best fit in the preceding illustration was found to be

$$y = .64x$$

when the origin was taken at the point of averages. In this equation y is expressed as a function of x; that is, x is taken to be the independent variable and y the dependent variable. The equation expresses the average variation in y (discount rates of commercial banks) corresponding to a change of one unit in x (discount rates of Federal Reserve Banks). This line of relationship corresponds precisely to a line of trend, which describes the average change in a given series accompanying a unit change in time. A line which thus describes the average relationship between two variables is termed a *line of regression*. Its equation is termed a *regression equation*, and the quantity $r\frac{\sigma_y}{\sigma_x}$ which gives the slope of such a line is called a *coefficient of regression*. The use of these terms dates back to early studies by Galton, dealing with the relation between the height of fathers and the height of sons. Sons, Galton found, deviated less on the average from the mean height of the race than their fathers. Whether the fathers were above or below the average, the sons tended to go back or *regress* towards the mean. He therefore termed the line which graphically described the average relationship between these two variables the *line of regression*. The term is now used generally, as indicated above, though the original meaning has no significance in most of its applications.

In any given case equations to two lines of regression may be computed. One is an expression of the average relationship between a dependent Y-variable and an independent X-variable; the other describes the relationship between a dependent X-variable and an independent Y-variable. The significance of the two may be indicated graphically.

Figure 70 is derived directly from the scatter diagram

presented in Fig. 69. The circle in each column represents the mean Y-value of all the items falling in that column. Thus in the first column there are 24 cases including all those with X-values falling between 3.75% and 4.25%.

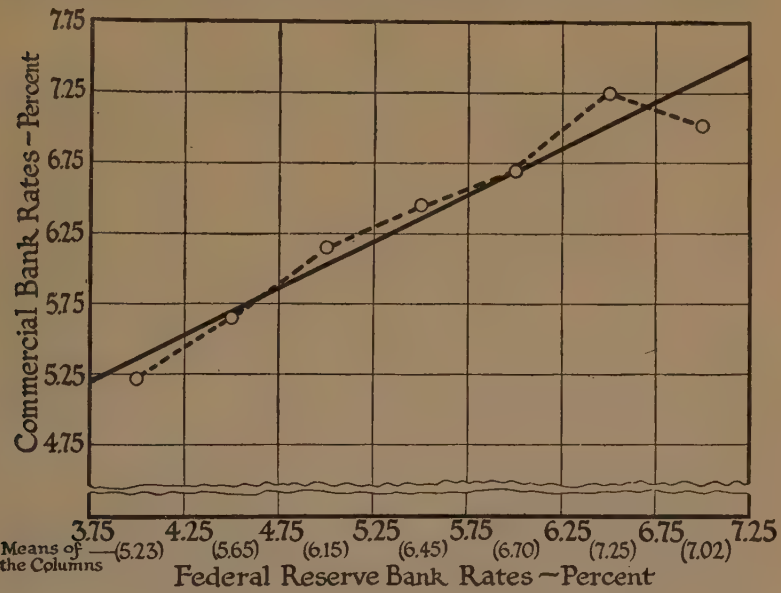


FIG. 70. — Showing the Relation between Discount Rates of Commercial Banks and Federal Reserve Bank Discount Rates. (The broken line connects the means of the columns and the straight line shows the average change in commercial bank rates corresponding to a unit change in Federal Reserve bank rates; i.e., it represents the regression of Y on X)

The Y-values vary, however, being distributed as shown in the following table.

TABLE 101

Computation of the Arithmetic Mean of an Array

Class-interval	*Mid-point* (m)	*Frequency* (f)	fm
5.75 – 6.24	6.0	1	6.0
5.25 – 5.74	5.5	11	60.5
4.75 – 5.24	5.0	10	50.0
4.25 – 4.74	4.5	2	9.0
		24	125.5

$$M = \frac{125.5}{24} = 5.23$$

Similar mean values are obtained for the other columns. These are plotted in Fig. 70, together with the line of regression of Y on X.

In Figure 70 the X-variable (Federal Reserve bank discount rates) is independent. As it increases from 4.0% to 4.5, 5.0, 5.5%, and so on, the average of commercial bank rates increases also. An average commercial bank rate of 5.23% was associated with an average Federal Reserve bank rate of 4%; an average commercial bank rate of 5.65% was associated with an average Federal Reserve bank rate of 4.5%, and so on. The slope of the straight line, which is the line of regression or the line of average relationship, measures the average increase in commercial bank rates corresponding to a unit increase in Federal Reserve bank rates.

It is possible to view the relationship between these two variables in another light. These questions arise: Given a certain commercial bank discount rate, what is the average Federal Reserve bank rate associated with it? And for a given change in commercial bank discount rates, what is the average change in the corresponding Federal Reserve bank rates? The commercial bank rate is now looked upon as independent, and the Federal Reserve rate as an associated dependent variable. These questions are answered by Fig. 71. The points marked by the small circles and connected by the broken line show the locations of the arithmetic means of the items falling in the various rows. Thus the three X-items in the first row have an average value of 4.17%. This is the average Federal Reserve bank discount rate associated with a commercial bank rate of 4.5%. The average Federal Reserve bank rate associated with a commercial bank rate of 5.0% is 4.35%, and so on. The straight line fitted to these points indicates the relationship between the two, its slope measuring the average

increase (or decrease) in Federal Reserve bank rates associated with a unit change in commercial bank rates.

This is the line of regression of X on Y. The general formula for the equation to this line is:

$$x = r\frac{\sigma_x}{\sigma_y}y$$

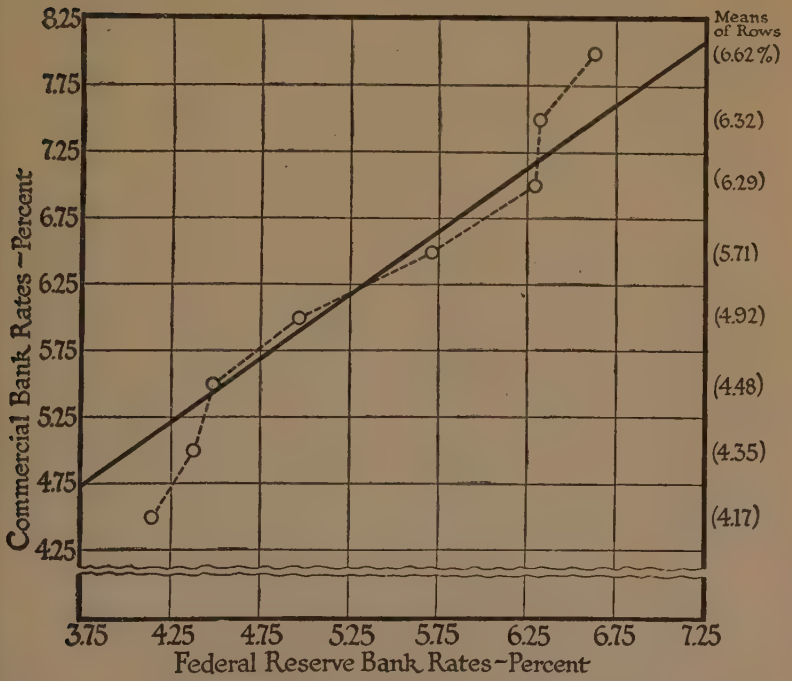


FIG. 71. — Showing the Relation between Federal Reserve Bank Discount Rates and the Discount Rates of Commercial Banks. (The broken line connects the means of the rows and the straight line shows the average change in Federal Reserve bank rates corresponding to a unit change in commercial bank rates; i.e., it represents the regression of X on Y)

Substituting the present values, we have

$$x = .82\frac{.8885}{.697}y$$

or

$$x = 1.045y$$

The factors in this equation, it will be seen, are the same as those entering into the formula for the line of regression of Y on X'. If r is equal to 1 the two lines coincide, and if, in addition, the two standard deviations are equal, the line of regression will bisect the angle formed by the axes. In any case, if the points be plotted on a chart scaled in units of the standard deviations, we have $y = rx$ and the slope of the line of regression is thus equal to the value of r.

The coefficient of regression is represented by the symbol b. In a simple correlation problem there are two such coefficients, representing the slopes of the two lines of regression. These are

$$b_{yx} = r\frac{\sigma_y}{\sigma_x}$$

$$b_{xy} = r\frac{\sigma_x}{\sigma_y}$$

(The subscripts indicate the relation between the two variables. The first subscript refers to the dependent variable in each case.)

The coefficient r appears in both formulas. This being so, it is clear that r may be computed from the regression coefficients. For

$$r = \sqrt{b_{yx} \cdot b_{xy}} = \sqrt{r\frac{\sigma_y}{\sigma_x} \cdot r\frac{\sigma_x}{\sigma_y}} = \sqrt{r^2}$$

[1] The formula $x = r\frac{\sigma_x}{\sigma_y}y$

may be reduced to $x = \frac{\Sigma(xy)}{\Sigma(y^2)}y$

This is the equation to a line fitted to the points plotted in Fig. 71 in such a way that the sum of the squares of the *horizontal deviations* is a minimum.

The formula

$$y = \frac{\Sigma(xy)}{\Sigma(x^2)}x$$

is the equation to the line for which the sum of the squares of the *vertical deviations* is a minimum. An understanding of this point may make clear the difference between the two lines of regression.

Thus if we know the slopes of the two lines of regression r may be determined. In the present example, using the exact values obtained,

$$r = \sqrt{.64248 \times 1.045} = .819$$

Use of the Equations of Regression

The two equations of regression given above

$$y = .64x$$

and

$$x = 1.045y$$

describe relations between deviations from the respective arithmetic means. That is, the origin is at the point of averages, and to use the equations we cannot use the original values of X and Y but must express them as deviations from their means. For example, we wish to determine the normal commercial bank rate associated with a Federal Reserve bank rate of 6%. The mean value of the X-variable (Federal Reserve bank rates) is 5.218%. A rate of 6% represents a deviation from the mean of +.782. Substituting this value in the first of the above equations, we have

$$\begin{aligned} y &= .64(+.782) \\ &= +.500 \end{aligned}$$

This is the average y-deviation associated with an x-deviation of +.782. To get the normal commercial bank rate associated with a Federal Reserve rate of 6% the value +.500% must be added to the mean commercial bank rate, 6.144%. The value we wish is thus 6.644%.

This calculation has been rather round-about because of the form of the equation of relationship. This equation can be put in more appropriate form for such computations.

Let

$\overline{X}$ = arithmetic mean of the X's

$\overline{Y}$ = arithmetic mean of the Y's.

Then

$$y = r\frac{\sigma_y}{\sigma_x}x$$

may be written

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X})$$

In this last equation X and Y represent the values of the variables on the original scales, and not as deviations from their respective means. In terms of the coördinate chart, it means shifting the origin from the point of averages to a point corresponding to zero on each of the original scales.

To illustrate the greater utility of the equation in this form, the equation

$$y = .64x$$

may be changed in the method indicated. It becomes

$$\begin{aligned} Y - 6.144 &= .64(X - 5.218) \\ &= .64X - 3.340 \\ Y &= .64X + 2.804 \end{aligned}$$

This is the equation with the origin so shifted that the original values may be employed directly. To determine the commercial bank rate normally associated with a Federal Reserve rate of 6% we may substitute the latter value in the equation just secured.

$$\begin{aligned} Y &= (.64 \times 6.0) + 2.804 \\ &= 6.644 \end{aligned}$$

Precisely the same results are secured as with the equation in the other form, but for many purposes it is preferable to have an equation in which the actual values may be inserted.

The equation

$$x = r\frac{\sigma_x}{\sigma_y}y$$

may be similarly changed to

$$X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$$

Summary

In the foregoing pages there have been presented two quite different methods of securing the values required in measuring the relationship between two variables. The steps in the two methods may be briefly summarized. The method of least squares is basic in both cases, but that term may appropriately be employed to describe the first method outlined, for the process of fitting the line is the first and fundamental step in that procedure.

I. The Least Squares Method.

A. Data to be handled as individual items.

1. Fit a straight line to the data by the method of least squares. A simple arrangement of the data in columns will permit the ready computation of the required values, $\Sigma(X)$, $\Sigma(Y)$, $\Sigma(X^2)$, $\Sigma(Y^2)$, $\Sigma(XY)$. The equation thus obtained describes the average relationship between the two variables.
2. Compute the standard error, S_y, from the formula

$$S_y^2 = \frac{\Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY)}{N}$$

S_y is a measure of the reliability of estimates based upon the equation of relationship, and is to be interpreted in the same way as is the standard deviation about an arithmetic mean.

3. Compute the coefficient of correlation, r, from the formula

$$r^2 = \frac{a\Sigma(Y) + b\Sigma(XY) - Nc_y^2}{\Sigma(Y^2) - Nc_y^2}$$

Give r the sign of the constant b in the equation of regression. This coefficient is an abstract measure of the degree of relationship between the two variables, in so far as this relationship may be described by a straight line.

4. If an equation describing the regression of X on Y (X being dependent) is desired, the proper values may be substituted in the two normal equations

$$\Sigma(X) = na + b\Sigma(Y)$$
$$\Sigma(XY) = a\Sigma(Y) + b\Sigma(Y^2)$$

The equation secured will be of the type

$$X = a + bY$$

The standard error, S_x, may be computed by making the appropriate changes in the formula as given for S_y. The value of r will be the same as in the preceding case, in which Y is dependent.

B. Data to be classified.

1. Select an appropriate class-interval and tabulate the items in the form of a correlation table.
2. Compute the necessary values for fitting a straight line to the data. In doing so, an arbitrary origin may be selected for each variable, and all values expressed in class-interval units. A re-arrangement in columnar form may facilitate the computation of the quantity

$$\Sigma(X'Y').$$

3. Compute the standard error, employing the formula given above.
4. Compute the coefficient of correlation from the formula given above.
5. If the above calculations were carried on in class-interval units the equation of average relationship and the standard error should now be expressed in terms of the original units of measurement. If an arbitrary origin was employed the equation should be corrected so that the variables relate to deviations from the true origin.

II. The Product-Moment Method.

A. Data to be handled as individual items.

1. Arrange the paired observations in parallel columns and compute the quantities $\Sigma(X)$, $\Sigma(Y)$, $\Sigma(X^2)$, $\Sigma(Y^2)$, $\Sigma(XY)$.
2. Divide these quantities throughout by N. For the first two of these quotients we may use the symbols c_x and c_y (i.e.

$$\frac{\Sigma(X)}{N} = c_x$$

and

$$\frac{\Sigma(Y)}{N} = c_y).$$

3. Compute the mean product from the formula

$$p = \frac{\Sigma(XY)}{N} - c_x c_y$$

4. Compute the two standard deviations from the formulas

$$\sigma_x = \sqrt{\frac{\Sigma(X^2)}{N} - c_x^2}$$

$$\sigma_y = \sqrt{\frac{\Sigma(Y^2)}{N} - c_y^2}$$

5. Compute the coefficient of correlation from the formula

$$r = \frac{p}{\sigma_x \sigma_y}$$

6. Determine the equations of regression by substituting the proper values in the formulas

$$y = r\frac{\sigma_y}{\sigma_x}x$$

$$x = r\frac{\sigma_x}{\sigma_y}y$$

(Note: For each of these equations the origin is at the point of averages.)

7. If desired, transfer the origin to zero on the two original scales by substituting the arithmetic means in the equations

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X})$$

$$X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$$

8. Compute the two standard errors from the formulas

$$S_y = \sigma_y\sqrt{1 - r^2}$$

$$S_x = \sigma_x\sqrt{1 - r^2}$$

B. Data to be classified.

1. Construct a correlation table as in I. B. above.
2. Select an assumed mean for each variable. Measure the deviations of the various items from the assumed means in class-interval units.
3. Compute c_x and c_y in class-interval units.
4. Compute σ_x and σ_y in class-interval units.
5. Compute $\Sigma(x'y')$ in class-interval units for each compartment of the correlation table. Total these figures to get $\Sigma(x'y')$ for the whole table.
6. Determine the value of the mean product in class-interval units from the formula

$$p = \frac{\Sigma(x'y')}{N} - c_x c_y$$

7. Compute r from the formula

$$r = \frac{p}{\sigma_x \sigma_y}$$

8. Reduce σ_x and σ_y to original units.
9. Determine the equations of regression by substituting the proper values in the formulas

$$y = r\frac{\sigma_y}{\sigma_x}x$$

and

$$x = r\frac{\sigma_x}{\sigma_y}y$$

10. If desired, transfer the origin to zero on the two original scales from the formulas

$$Y - \overline{Y} = r\frac{\sigma_y}{\sigma_x}(X - \overline{X})$$

$$X - \overline{X} = r\frac{\sigma_x}{\sigma_y}(Y - \overline{Y})$$

11. Compute the two standard errors from the formulas

$$S_y = \sigma_y\sqrt{1 - r^2}$$

$$S_x = \sigma_x\sqrt{1 - r^2}$$

It is advisable, in all cases, to construct scatter diagrams and to plot the lines of regression thereon. It is generally

possible to derive from such diagrams a truer idea of the relations involved, and of the adequacy of the methods employed, than may be obtained from a study of the figures alone.

Limitations

A question naturally arises as to the degree of generality attaching to the measures of relationship described in the preceding pages. Are they limited to certain types of distributions, or may they be employed as absolutely general and universally valid measures?

As we have seen, the standard deviation has a precise and definite meaning with respect to distributions following the normal law. Having values of the mean and of the standard deviation, we know, in such cases, the exact percentage of observations which will fall within any stated limits. If the distribution departs from the normal type the standard deviation is still a useful measure, but it cannot be interpreted in the same exact sense. Bearing this in mind, the formula

$$r^2 = 1 - \frac{S_y^2}{\sigma_y^2}$$

may be considered.

When the distribution of the original values of the dependent variable about their mean is normal and the distribution about the least squares line is normal, both S_y and σ_y have specific and exact meanings, and it is perfectly legitimate to compute such a measure as r, based upon the relation of one to the other. Departures from normality in either case reduce the significance of this comparison. But we have seen that the standard deviation remains a useful measure even though the departure from the normal type be fairly pronounced, though in the latter case, it lacks the precise significance attaching to it in a normal distribution. In the same way the standard error and the coefficient of correlation may be computed and utilized, even when all the requirements of normality are not met.

Care must be taken in their interpretation in such cases, however. It must be clearly recognized that these measures have their full significance only in cases where the original distribution of the dependent variable and the distribution about the least squares line are both normal, or approximately so.

A simple example may make clear the effect upon the value of the coefficient of correlation of an extreme departure from a normal distribution. In the following table are listed certain selected figures taken from the 1919 Census of Manufactures, for the State of New York.

TABLE 102

Wage Earners in Factories and Value of Products, 1919, in Eleven Cities in the State of New York

City	*Number of Wage Earners (in thousands)* (*X*)	*Total Value of Products (in millions of dollars)* (*Y*)
Batavia	2.2	9
Beacon	2.2	10
Corning	3.5	11
Geneva	2.5	10
Glens Falls	2.8	12
Ithaca	1.7	10
Middletown	2.2	10
Peekskill	2.1	11
Rensselaer	1.4	10
Tonawanda	1.8	16
New York City	638.8	5,261

When the first ten of these cities, in the order listed, are treated as a group, the following values are secured:

$$\sigma_y = 1.8682$$
$$S_y = 1.8669$$
$$r = -.034$$

The ten points and the line of regression are plotted in Fig. 72. (No general significance is to be attached to the

above coefficient of correlation, for the cities were selected for the purpose of illustrating a particular point.)

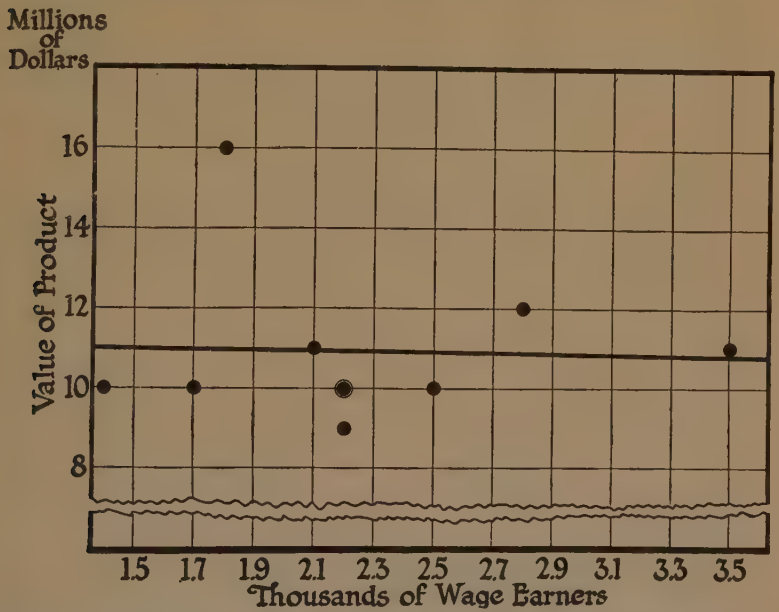


FIG. 72. — Showing the Relation between Number of Wage Earners in Factories and Value of Products in Ten Selected Cities in the State of New York

When New York City is included in the group, the values secured for the sample of eleven cities are

$$\sigma_y = 1509.3$$
$$S_y = 7.53$$
$$r = +.999988$$

The eleven points and the line of regression are plotted in Fig. 73.

The reason for the markedly different results is obvious. When the one very large city is included with the ten small cities the standard deviations of both variables are greatly increased. That of the Y-variable (value of products) is increased from 1.8682 to 1509.3. But S_y, the measure of the scatter about the fitted line, undergoes no such pro-

nounced change in value. For the ten cities it is 1.8669; for the eleven cities 7.53. This is due to the fact that the one exceptional case is given such a great weight, in fitting by the method of least squares, that the fitted line must

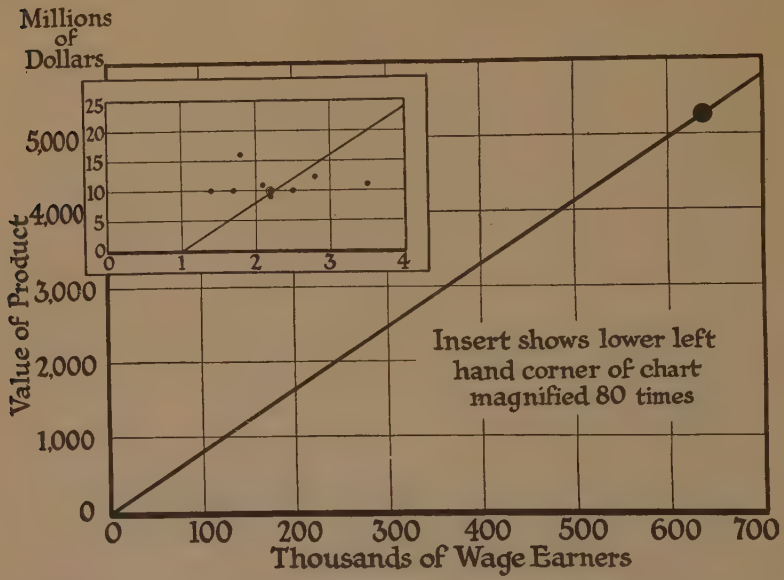


FIG. 73. — Showing the Relation between Number of Wage Earners in Factories and Value of Products in Eleven Selected Cities in the State of New York

pass through or very near the point representing this observation. Accordingly, S is always affected less than σ by a single very exceptional case. Since the value of r depends upon the relationship

$$r = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}}$$

the presence of such a case always tends to increase the value of the measure of correlation. The introduction of the one exceptional case in the above example changes a correlation coefficient of virtually zero to one of unity. The result, of course, is meaningless.

While this example represents an extreme instance, the

same distortion will be felt, to a greater or less degree, whenever there is a departure from a normal distribution. In practice the various measures of relationship cannot be restricted to perfectly normal distributions, but they must be interpreted with care when there is reason to believe that such disturbing influences are present.

REFERENCES

Bowley, Arthur L. *Elements of Statistics* (350–397).

Brunt, David. *The Combination of Observations* (148–170).

Chaddock, R. E. *Principles and Methods of Statistics* (Chap. XII).

Elderton, W. P. *Frequency Curves and Correlation* (106–124).

Galton, Francis. *Correlations and their Measurement.* Proceedings of the Royal Society. Vol. XLV, 1888 (136–145).

Jones, D. C. *A First Course in Statistics* (102–131).

Kelley, Truman L. *Statistical Method* (151–195).

Moore, H. L. *Forecasting the Yield and the Price of Cotton* (12–51).

Pearl, Raymond. *Medical Biometry and Statistics* (292–318).

Pearson, Karl. *Regression, Heredity and Panmixia.* Phil. Transactions, Royal Society. Series A. Vol. CLXXXVII, 1896 (253–318).

Rietz, H. L. and Crathorne, A. R. *Simple Correlation* (In Rietz, H. L. *Handbook of Mathematical Statistics*, 120–129).

Rugg, H. O. *Statistical Methods Applied to Education* (233–307).

Whitaker, E. T. and Robinson, G. *The Calculus of Observations* (317–336).

Yule, G. U. *An Introduction to the Theory of Statistics* (157–209).

CHAPTER XI

THE MEASUREMENT OF RELATIONSHIP BETWEEN TIME SERIES

The methods of measuring correlation which have been described in the preceding chapter were devised originally for the analysis of non-historical data, that is, for the treatment of *frequency* series rather than *time* series. The measurement of correlation between series in time presents certain distinctive problems which require separate treatment.

We have seen that such series are affected by various forces, which have been classified as the secular trend, cyclical and seasonal fluctuations and accidental variations, and methods have been described by means of which the effects of these various forces may be isolated. This breaking up of a series into its component parts for separate study is essential in attempting to correlate series in time, for spurious and quite misleading results will be secured if this is not done. The problem of correlation is that of securing a precise measure of the degree of relationship between variable quantities. But each series in time represents the combination of a number of variables and, so far as possible, each should be treated separately in correlating such series.

The relationship between two time series as, for example, interest rates and bond prices, may be studied with respect to any or all of the following components:

a. Secular trend.

b. Cyclical fluctuations.

c. Seasonal fluctuations.

d. Changes from one time unit to the next (*e.g.*, week to week, month to month, or year to year).

Such relationships may be studied, first, through the comparison of graphs, and much may be learned by this simple process. The similarity or dissimilarity of secular trends, and the general relation between cyclical movements may be determined by a study of such graphs. For more accurate comparison the coefficient of correlation may be used, but when it is so employed it is particularly important that the precise nature of its employment and the exact significance of the results be understood.

For the comparison of secular trends the coefficient of correlation would never be employed. The mere fact that two series have the same secular trend is no indication of a relationship of interdependence; a coefficient of correlation based upon the trend values would be meaningless. Moreover, much simpler methods are available for comparing trends.

For the same reason a coefficient of correlation should not be based upon the original absolute values of two series in time, except in the rather rare case in which neither series is marked by a definite secular trend. The computation of r, when dealing with ordinary statistical data, involves measuring the deviations of all the items from their respective arithmetic means, and securing the sum of the products of the paired deviations. When deviations of like sign are paired throughout r will have a positive value; when deviations of unlike signs are paired throughout r will have a high negative value. The presence of a pronounced upward or downward secular trend makes it impossible to secure significant values for r by the employment of this method. For example, the relationship between two series, as automobile production and the price of bacon between the years 1900 and 1920, might be measured. The secular trend is markedly upward in both cases. When the deviations of the annual figures are measured from the arithmetic means of the two series, the paired items for the earlier years will be negative, for the later years positive.

A fairly high positive value for r would be secured, were the computation carried through on this basis. This value would be quite misleading, for no real relationship can be expected in this case. The coefficient of correlation in such a case would measure, primarily, the relation between the two secular trends.

This coefficient might conceivably be employed to determine the similarity between seasonal fluctuations in two series, but its utility for this purpose may be questioned. Here again other and simpler methods are available.

In practice, therefore, the device of correlation should be employed neither to measure the relation between secular trends nor between seasonal movements. Its use is confined to comparisons of two or more series with respect to cyclical fluctuations and with respect to the short time changes from month to month or year to year. And, if valid measures of correlation are to be secured in making such comparisons, the effects of other forces which are not being studied must be eliminated, in so far as this is possible. The actual work of correlation must be preceded by a sifting process designed to remove all such irrelevant material. Unless the data are thus "distilled" the interpretation of the resulting coefficients will be difficult.

The Measurement of Correlation Between Cyclical Fluctuations

In an earlier chapter we have dealt with methods by which the effects of certain of the factors affecting time series might be measured and eliminated. The spurious correlation due to secular trend may be avoided by measuring the deviations of the observations not from the respective arithmetic averages but from the lines of secular trend of the two series. These variations, the deviations from normal, are the significant values if our interest centers

in the cycles. If annual values are employed the problem of eliminating seasonal fluctuations is not faced.

To illustrate this method of measuring the relationship between series in time we may undertake to determine whether there is any connection between cyclical fluctuations in cotton production and in cotton prices. Figures for crop years are to be employed, for the period 1900–01 to 1922–23.

Cotton prices require some correction before correlation is attempted. The raw figures with which the investigation

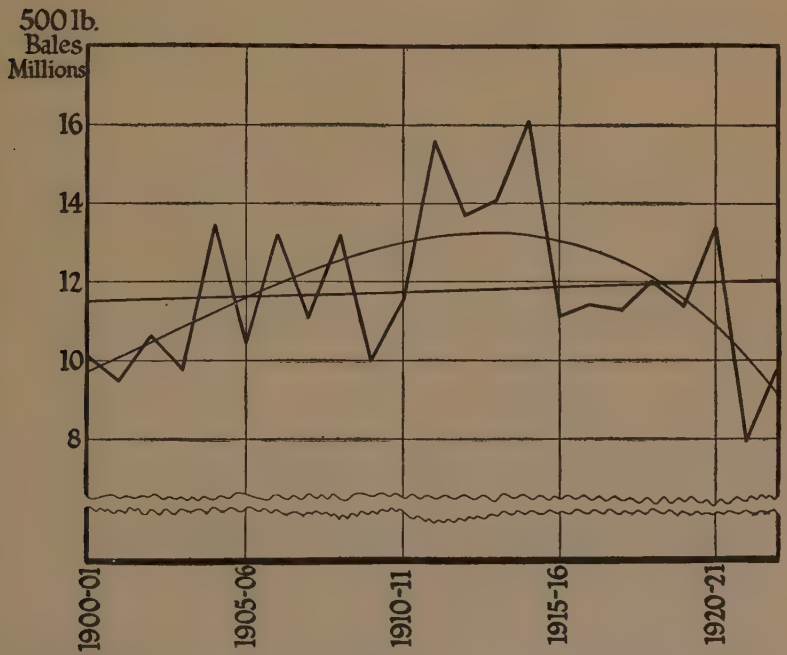


Fig. 74. — Cotton Production in the United States, Crop Years 1900–01 to 1922–23, with Lines of Trend

starts are average spot prices at New York for middling upland cotton, at wholesale, from September to May of each crop year. But such prices reflect not only the effects of varying conditions in the cotton market, but also changes

in the general level of prices. To eliminate the effect of this factor the original prices are deflated by Bradstreet's price index, as computed for the September-May period in each crop year. For this purpose Bradstreet's index has been reduced to relative terms, with the average for the crop year 1913–14 equal to 100. The original figures for the two series to be correlated, together with the corrected price figures, are given in Table 103.

TABLE 103

Cotton Production and Cotton Prices, 1900–1923

(1) *Crop Year*	(2) *Cotton Production in United States, excluding linters (in thousands of bales)*	(3) *Cotton Prices. Average of spot prices in N.Y. for middling upland cotton, Sept. to May (in cents per pound)*	(4) *Bradstreet's Price Index Average, Sept. to May (1913-14=100)*	(5) *Cotton Prices, deflated (in cents per pound)*
1900–01	10,123	9.58	84.8	11.30
1901–02	9,510	8.64	86.2	10.02
1902–03	10,631	9.50	90.0	10.56
1903–04	9,851	13.20	88.6	14.90
1904–05	13,438	8.69	89.3	9.73
1905–06	10,575	11.40	92.3	12.35
1906–07	13,274	10.97	98.8	11.10
1907–08	11,107	11.41	93.2	12.24
1908–09	13,242	9.81	91.3	10.74
1909–10	10,005	14.62	100.6	14.53
1910–11	11,609	14.80	97.8	15.13
1911–12	15,693	10.34	100.0	10.34
1912–13	13,703	12.35	104.8	11.78
1913–14	14,156	13.40	100.0	13.40
1914–15	16,135	8.63	105.2	8.20
1915–16	11,192	12.04	121.2	9.93
1916–17	11,450	18.29	151.0	12.11
1917–18	11,302	29.96	197.9	15.14
1918–19	12,041	30.06	203.1	14.80
1919–20	11,421	38.63	226.3	17.07
1920–21	13,440	16.90	152.9	11.05
1921–22	7,954	18.67	127.2	14.68
1922–23	9,762	26.26	149.7	17.54

These data are plotted in Figures 74 and 75. Two lines of trend which were fitted to each series are shown on each chart.

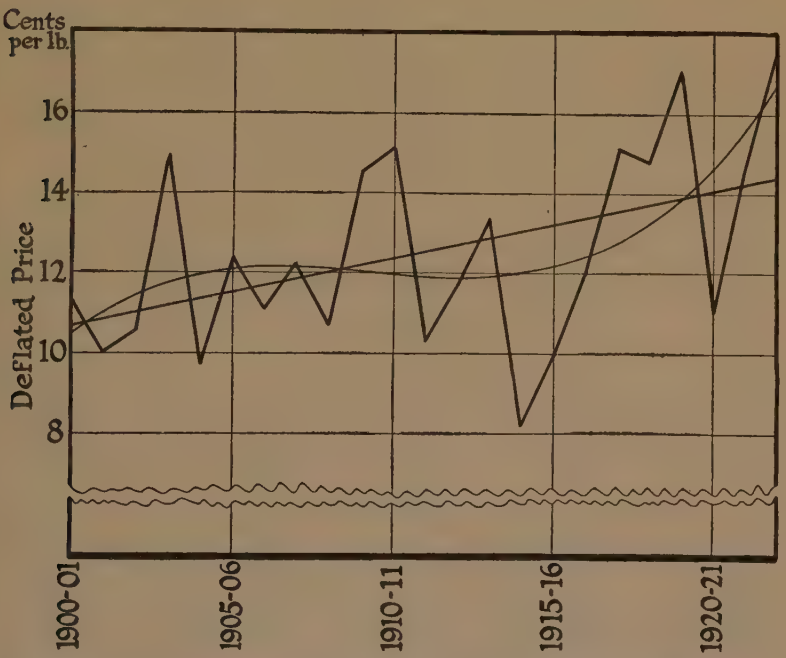


FIG. 75. — Prices of Middling Upland Cotton in New York, Crop Years 1900–01 to 1922–23, with Lines of Trend. (Figures relate to average annual prices, during the crop years, deflated by Bradstreet's index of wholesale prices)

The deviation of each annual item from the secular trend of the given series is now to be measured, and the coefficient of correlation between these deviations is to be calculated. Deviations from the two third degree parabolas are first correlated.[1] The computations appear in Table 104.

[1] It is not, of course, essential that the trend lines be of the same type for the two series being correlated.

TABLE 104

Computation of the Coefficient of Correlation, Cotton Production and Cotton Prices

(1) Year	(2) Deviation of Cotton Production from parabolic line of trend (in thousands of bales) x	(3) Deviation of deflated Cotton Price from parabolic line of trend (in cents per pound) y	(4) x^2	(5) y^2	(6) xy
1900–01	+371	+.79	137,641	.6241	+293.09
1901–02	−603	−1.05	363,609	1.1025	+633.15
1902–03	+145	−.93	21,025	.8649	−134.85
1903–04	−1,012	+3.11	1,024,144	9.6721	−3,147.32
1904–05	+2,200	−2.26	4,840,000	5.1076	−4,972.00
1905–06	−1,027	+.24	1,054,729	.0576	−246.48
1906–07	+1,325	−1.06	1,755,625	1.1236	−1,404.50
1907–08	−1,164	+.08	1,354,896	.0064	−93.12
1908–09	+681	−1.38	463,761	1.9044	−939.78
1909–10	−2,806	+2.47	7,873,636	6.1009	−6,930.82
1910–11	−1,405	+3.14	1,974,025	9.8596	−4,411.70
1911–12	+2,530	−1.60	6,400,900	2.5600	−4,048.00
1912–13	+452	−.13	204,304	.0169	−58.76
1913–14	+887	+1.47	786,769	2.1609	+1,303.89
1914–15	+2,923	−3.81	8,543,929	14.5161	−11,136.63
1915–16	−1,878	−2.24	3,526,884	5.0176	+4,206.72
1916–17	−1,388	−.30	1,926 544	.0900	+416.40
1917–18	−1,206	+2.37	1,454,436	5.6169	−2,858.22
1918–19	−31	+1.56	961	2.4336	−48.36
1919–20	−102	+3.21	10,404	10.3041	−327.42
1920–21	+2,586	−3.59	6,687,396	12.8881	−9,283.74
1921–22	−2,104	−.91	4,426,816	.8281	+1,914.64
1922–23	+636	+.82	404,496	.6724	+521.52
Totals	0	0	55,236,930	93.5284	−40,752.29

$$\sigma_x = \sqrt{\frac{55,236,930}{23}} = 1,550$$

$$\sigma_y = \sqrt{\frac{93.5284}{23}} = 2.017$$

$$r = \frac{\Sigma(xy)}{N\sigma_x\sigma_y} = \frac{-40,752.29}{23 \times 1,550 \times 2.017}$$

$$= -.567$$

This value of $-.567$ for the coefficient indicates a fair degree of negative correlation between deviations of cotton production in the United States from the line of trend and the corresponding deviations of cotton prices in New York, during the period covered. A somewhat higher value could undoubtedly have been secured had the rather abnormal war years been omitted, but there are some objections to such an omission.

From the values already computed we may derive an equation for estimating the variation in cotton price associated with a given variation in production. This regressive equation, as we have seen, is of the type

$$y = r\frac{\sigma_y}{\sigma_x}x$$

In the present case y and x refer to deviations from the parabolic lines of trend. Substituting the given values we have

$$y = -.567\frac{2.017}{1{,}550}x$$

$$y = -.0007378x$$

This equation means that, on the average, a unit deviation of cotton production (x) above the line of trend was accompanied by a deviation of .0007 units in cotton prices (y) below the line of trend. The unit employed in the production figures was 1000 bales, in the deflated price figures, one cent. In the interpretation of the equation it may be simpler to use an x-unit of one million bales, making the equation of regression

$$y = -.7378x$$

Thus a cotton crop one million bales above normal was accompanied by prices about three quarters of a cent per pound below normal (with reference always to deflated prices). This was the average relationship during the period 1900–1923. It did not hold in all cases, as is shown by the fact that r has a value of but $-.567$. If this, or

a similar law, held perfectly, r would have a value of -1.

The value of S, which measures the scatter about the line of regression, may be computed from the formula

$$S_y = \sigma_y \sqrt{1 - r^2}$$

In the present case, S_y has a value of 1.66 cents. The significance of this measure has been explained in an earlier section.

(It should be emphasized that the use of the above equation for estimating future prices is dependent upon the validity of projecting the two lines of secular trend.)

In the preceding analysis deviations were measured in absolute units, and the results could be interpreted only in terms of absolute units, bales of cotton and cents per pound. For certain purposes it might have been more convenient to correlate percentage deviations from the two lines of trend, in which case the standard deviations and the equation of regression would have been expressed in these terms. The procedure, in this respect, will generally depend upon the use to which the results are to be put.

It is obvious that in the above problem there is an arbitrary element which was not present in the correlation problems previously studied. The deviations are measured from lines of trend, not from the arithmetic means, and these lines of trend are arbitrarily selected. The use of different lines of trend might give quite different results. In the above example the lines of trend were both third degree parabolas. We may, with perhaps equal reason, assume that the underlying trend is best described by a straight line in each case. When such lines are fitted, and the deviations from these linear trends correlated, a value of $-.608$ is secured for the coefficient of correlation, with S_y having a value of 1.72 cents. (These results may appear inconsistent with those secured when deviations from the parabolic lines of trend are correlated, for both S_y and r are greater in the case in which linear trends are

assumed. The explanation is found in the fact that σ_y is also greater in the case of the linear trend; the value of r, of course, depends upon the relation between S_y and σ_y.)

Difficulties in the Correlation of Time Series

Which of these coefficients gives the true measure of the relationship between cyclical fluctuations in the two series? This question, unfortunately, cannot be answered definitively, and in this fact is found one of the chief difficulties faced in the correlation of time series. If we could say with certainty which of the two curves we have fitted to each series is the best measure of trend the question could be answered, but there is no final objective test to employ. It would appear from inspection that the curve of higher degree represents the trend rather better within the limits of the data in each case, but that is not capable of proof. Certainly it does not follow that the lines of trend which give the highest coefficient of correlation are the best. A high value for the coefficient may mean only that the two lines of trend are bad fits, both erring in the same way.[1]

The presence of this arbitrary element in the correlation of deviations from lines of secular trend detracts somewhat from the confidence that may be placed in the results. With the selection of different lines of trend a number of quite different results may be secured for the coefficient of correlation between cyclical fluctuations in two series. The critical problem in such a case lies not in the mechanical process of correlation, but in the choice of an appropriate line of trend for each series. If, by the tests of inspection and of correspondence with such external evidence as may be available, it appears that the curve selected accurately represents the trend in each of the series correlated, the coefficient may be accepted as significant. But, in the

[1] *Cf.* W. M. Persons, "Correlation of Time Series," *Journal of the American Statistical Association*, June, 1923, 725.

interpretation and use of the results, the presence of this element of personal judgment in the preliminary calculations must not be forgotten. This applies with particular force if the study aims to establish a functional relationship between cyclical fluctuations in the two series, and if an estimating (or regression) equation is to be based upon the results.

The Coefficient of Correlation and the Measurement of Time Sequence

In the correlation of cotton production and cotton prices the object was to measure as accurately as possible the effect of variations in cotton production upon cotton prices. An equation was secured which described this relation when deviations were measured from the particular lines of trend employed. Cotton prices were considered to be a function of cotton production, and the object of the study was to measure this functional relationship. We seek, in such cases, to determine the degree to which cycles in one series depend upon or reflect cycles in a related series, assuming some functional relationship between them. This is essentially the problem described in introducing the subject of correlation, and generally constitutes the major problem in studying the relation between series of any type.

But a second and somewhat different problem may be faced in certain studies of time series. Assuming that two such series are marked by definite cycles, it is of interest to determine whether the cycles coincide in time, or whether cycles in one series consistently precede or lag behind cycles in the other. The coefficient of correlation has been found very useful in determining the degree of "lead" or "lag" in such cases. This problem is that of determining merely *temporal relationship*, as opposed to the *functional relationship* which is ordinarily to be measured.

The Relation between Stock Price Cycles and Cycles of Business Activity

To illustrate the solution of a problem of this latter type we may undertake to determine the relation, in time, between cyclical movements in industrial stock prices and in general business activity, as measured by the composite index compiled by the American Telephone and Telegraph Company. The monthly values of this index for the period 1877–1923 have been presented in an earlier section. The figures relating to stock prices are given in the following table:

Table 105

Cycles in Industrial Stock Prices, 1903–1923[1]

(Figures relate to deviations from normal in units of the standard deviation.)

	1903	1904	1905	1906	1907	1908	1909	1910	1911	1912	1913
January	−.2	−2.0	−.1	+2.3	+1.6	−1.3	+.5	+1.0	−.1	−.4	−.2
February	−.1	−2.1	+.2	+2.2	+1.4	−1.5	+.3	+.5	+.1	−.4	−.5
March	−.3	−2.1	+.6	+1.9	+.6	−1.1	+.3	+.8	−.1	−.1	−.7
April	−.5	−2.0	+.7	+1.7	+.6	−.8	+.5	+.5	−.1	+.3	−.6
May	−.5	−2.1	+.2	+1.4	+.4	−.5	+.8	+4	0	+.2	−.7
June	−.9	−2.1	+.3	+1.5	+.2	−.5	+.9	+.1	+.1	+.2	−1.1
July	−1.4	−1.8	+.7	+1.3	+.3	−.2	+1.1	−.4	+.1	+.2	−.9
August	−1.7	−1.6	+.8	+1.7	−.3	+.3	+1.4	−.3	−.3	+.3	−.7
September	−1.9	−1.3	+.7	+1.7	−.5	+.1	+1.4	−.3	−.7	+.4	−.5
October	−2.3	−.9	+.8	+1.7	−1.3	+.2	+1.4	0	−.7	+.3	−.8
November	−2.4	−.3	+1.1	+1.7	−1.9	+.5	+1.4	+.1	−.5	+.2	−.9
December	−2.1	−.1	+1.8	+1.7	−1.6	+.5	+1.3	−.2	−.4	0	−.9

	1914	1915	1916	1917	1918	1919	1920	1921	1922	1923	
January	−.7	−2.42	+.42	+.61	−1.00	−.56	+1.16	−1.04	−.61	+.66	
February	−.6	−2.47	+.33	+.15	−.68	−.53	+.40	−1.01	−.37	+.95	
March	−.6	−2.30	+.31	+.41	−.80	−.13	+.78	−1.02	−.13	+1.14	
April	−.8	−1.69	+.05	+.36	−.86	+.14	+.88	−.92	+.19	+.93	
May	−.8	−1.72	+.08	+.32	−.61	+.77	+.17	−.88	+.34	+.53	
June	−.7	−1.54	+.11	+.59	−.63	+1.22	+.17	−1.49	+.32	+.35	
July	−1.1	−1.28	−.04	+.29	−.59	+1.56	+.11	−1.54	+.45	+.01	
August	*	−.74	+.15	−.03	−.54	+1.01	−.27	−1.66	+.69	+.13	
September	*	−.27	+.62	−.41	−.50	+1.38	−.16	−1.41	+.81	+.09	
October	*	+.25	+.97	−.75	−.22	+1.86	−.31	−1.32	+.84	−.11	
November	*	+.40	+1.40	−1.32	−.35	+1.62	−.80	−.97	+.51	+.14	
December	−2.54	+.58	+.70	−1.41	−.50	+1.24	−1.28	−.69	+.64	+.37	

* Stock Exchange closed.

[1] These figures, the results of analyses made by W. M. Persons, are from the *Review of Economic Statistics*, published by the Harvard Committee on Economic Research. For the period January, 1903, to July, 1914, they are based upon the average price of 12 industrial stocks; for the period since December, 1914, they are based upon the average price of 20 industrial stocks (the Dow-Jones index).

The data of the two series are plotted in Fig. 76. From a comparison of the two curves in this chart it is clear that there is some relation between the movements in the two series, but such a comparison affords no basis for a definite conclusion. Our object is to determine whether the cycles in the two series are exactly synchronous and, if they are not, to measure the average time interval by which cycles in one series precede the cycles in another. The significance of such studies in the analysis of the business cycle is obvious.

For the purpose of this investigation data for the period from January, 1903, to June, 1914, may be employed. The war years are omitted because the two series, during these years, were affected by the abnormal conditions prevailing.

A coefficient of correlation is first computed for concurrent items. A value of $+.55$ is secured. Next, the data are correlated with industrial stock prices preceding general business by one month. That is, the January, 1903, figure for stock prices is multiplied by the February, 1903, index of general business; the February stock price is multiplied by the March business index, etc. This process is carried through for the entire period from January, 1903, to June, 1914. Only 137 monthly values are used in this computation, as compared with 138 in the preceding case, for the January, 1903, business index and the June, 1914, stock price figure do not enter into the calculations. Accordingly the values c_x and c_y (the two corrections to be applied because the origin does not coincide with the two averages) and the two standard deviations will be slightly different. These corrections may be readily made. The coefficient of correlation secured from these computations has a value of $+.65$. The same operation is repeated with other pairings of the two variables. The results are summarized below.

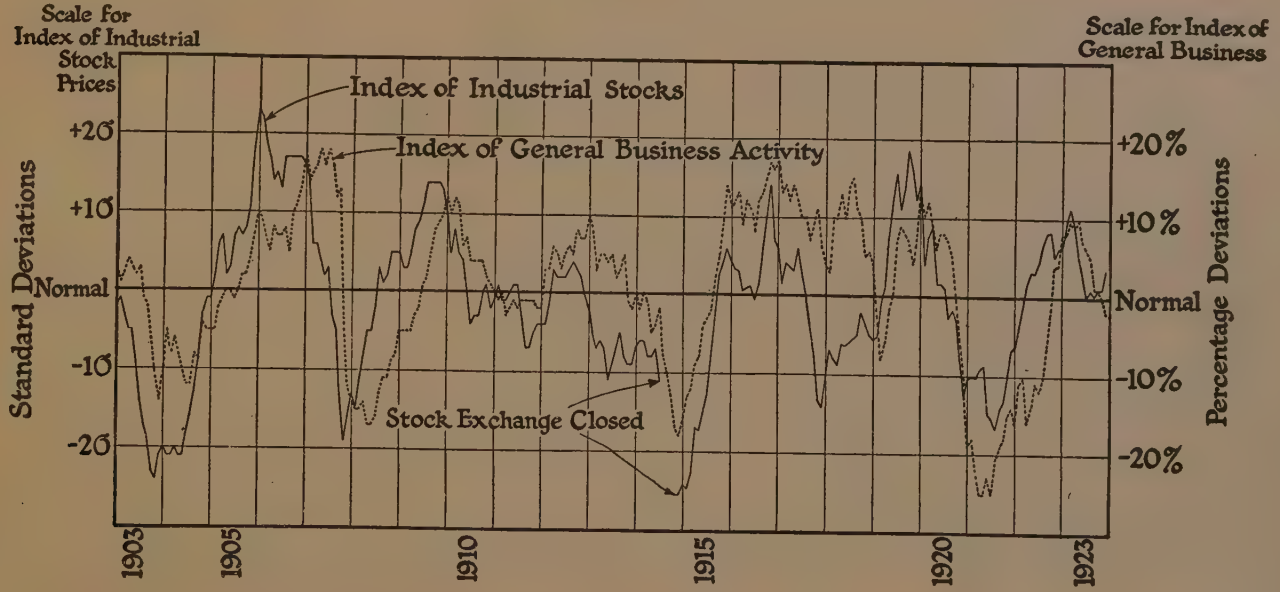


FIG. 76. — Comparison of the Cyclical Fluctuations in Industrial Stock Prices and in General Business Activity, 1903–23

TABLE 106

Coefficients of Correlation between Industrial Stock Prices and an Index of General Business Activity

(Based upon data for the period 1903–1914.)

	Coefficient of Correlation
Stock prices concurrent with business index	+ .55
Stock prices preceding business index by one month	+ .65
" " " " " " 2 months	+ .70
" " " " " " 3 "	+ .73
" " " " " " 4 "	+ .76
" " " " " " 5 "	+ .76
" " " " " " 6 "	+ .76
" " " " " " 7 "	+ .74
" " " " " " 8 "	+ .71
" " " " " " 9 "	+ .67
" " " " " " 10 "	+ .61
" " " " " " 11 "	+ .54

These figures are plotted in Fig. 77.

The coefficients increase to a maximum value of +.76 which is secured with stock prices preceding general business

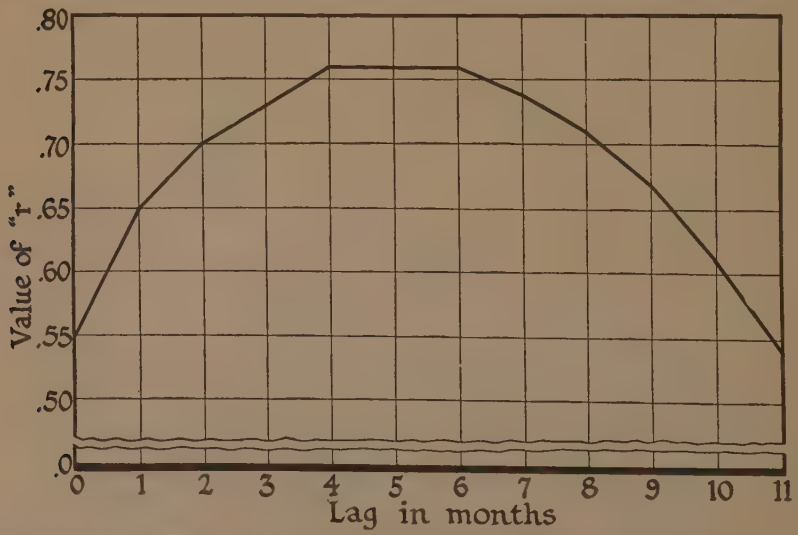


FIG. 77. — Coefficients of Correlation between Index of Industrial Stock Prices and Index of Business Activity, 1903–14, Showing the Results Secured with Different Pairings

(In all pairings except that of concurrent items the business activity index follows the stock price index)

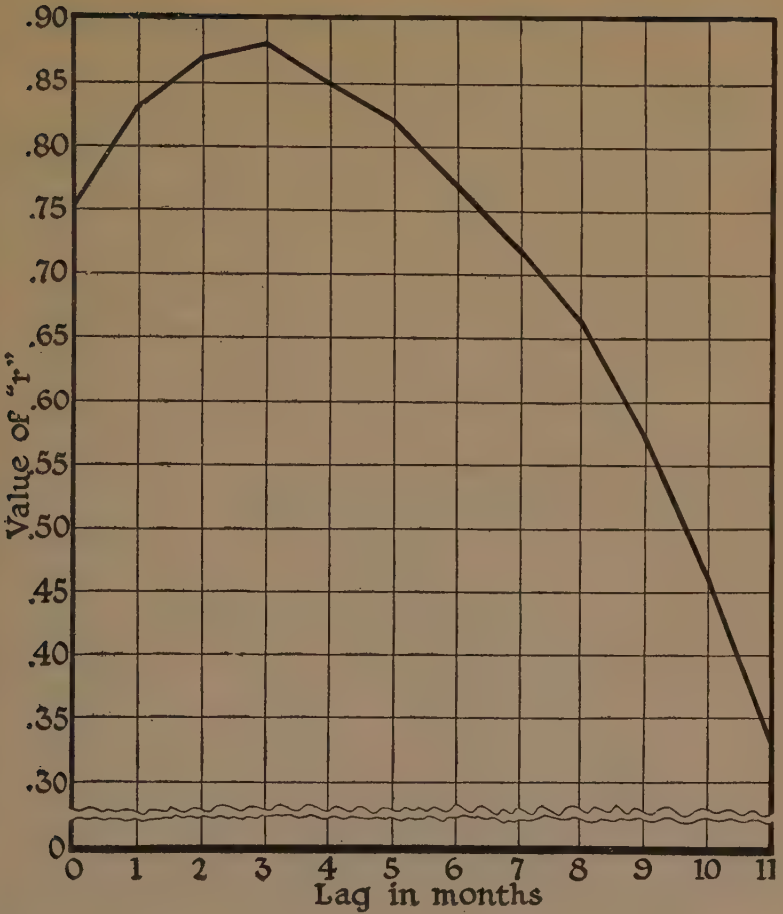


FIG. 78. — Coefficients of Correlation between Index of Industrial Stock Prices and Index of Business Activity, 1919–23, Showing the Results Secured with Different Pairings

(In all pairings except that of concurrent items the business activity index follows the stock price index)

by 4, 5 and 6 months. The stability of the coefficients with the period of "lead" varying from 3 to 7 months indicates that there is no one specific interval, within the limits thus indicated, between the cyclical movements of these two series. From the results here given it would

appear that five months is the average interval by which stock prices precede the general business index, but this is not sharply marked off as a constant relationship.

When the same figures are correlated for the post war period, from January, 1919, to December, 1923, somewhat different results are secured. These are presented below; they are plotted in Fig. 78.

TABLE 107

Coefficients of Correlation between Industrial Stock Prices and an Index of General Business Activity

(Based upon data for the period 1919–1923.)

	Coefficient of Correlation
Stock prices concurrent with business index	+ .75
Stock prices preceding business index by one month	+ .83
" " " " " " 2 months	+ .87
" " " " " " 3 "	+ .88
" " " " " " 4 "	+ .85
" " " " " " 5 "	+ .82
" " " " " " 6 "	+ .77
" " " " " " 7 "	+ .72
" " " " " " 8 "	+ .66
" " " " " " 9 "	+ .57
" " " " " " 10 "	+ .46
" " " " " " 11 "	+ .33

The period covered in this latter study is so short that no final conclusions may be drawn. The results indicate that since the war the movements of general business have followed more closely behind stock price movements than during pre-war days. A maximum correlation is secured when stock prices precede the business index by three months.

THE USE OF THE MOVING AVERAGE IN CORRELATING CYCLES IN TIME SERIES

The preceding discussion has dealt only with cycles as measured from mathematically fitted lines of trend. But trend may be measured, as we have seen, by lines based

upon moving averages, and the cyclical deviations from such lines may be correlated in precisely the same way as deviations from other lines of trend. The arithmetic mean of the deviations from such moving averages will not necessarily be zero, as in the case of deviations measured from lines fitted by the method of least squares, and a corresponding correction must be made in correlating such figures.

Moving averages are subject to the same criticism as are mathematical lines of trend. There can be no certainty that deviations from lines of trend based upon moving averages represent the effects of cyclical causes solely. The result in a given case depends upon the period of the moving average employed, and there is no perfect criterion by which to determine the best measure of trend. Significant and useful coefficients may be computed when deviations are measured from moving averages, but the presence of an arbitrary element in the work must be recognized and the results applied with corresponding reservations.

The Correlation of Short Term Fluctuations

In describing the variable factors that constitute component elements of the values of a series in time, it was pointed out that the coefficient of correlation would not generally be employed in comparing either the secular trends or the seasonal fluctuations of two series. It may be used to advantage in measuring either *functional* or *temporal* relations between cyclical fluctuations, provided that the effects of the other variables have been, so far as possible, eliminated. The coefficient of correlation and the measures which are employed in conjunction with it have a further use in dealing with time series. They may be used to measure the relation between short term changes in two series, changes from year to year, month to month, or even from

week to week or day to day, if desired. This problem is distinct from that studied in the preceding section and in the interpretation of the results the two should not be confused.

TABLE 108

Computation of Coefficient of Correlation between Cotton Production and Cotton Prices, 1901–1923

(Based upon first differences.)

(1) *Crop Year*	(2) *Difference between production in given year and production in preceding year (in millions of bales)* X	(3) *Difference between price in given year and price in preceding year (in cents per pound, deflated)* Y	(4) X^2	(5) Y^2	(6) XY
1901–02	−.613	−1.28	.375769	1.6384	+.78464
1902–03	+1.121	+.54	1.256641	.2916	+.60534
1903–04	−.780	+4.34	.608400	18.8356	−3.38520
1904–05	+ 3.587	−5.17	12.866569	26.7289	−18.54479
1905–06	−2.863	+2.62	8.196769	6.8644	−7.50106
1906–07	+2.699	−1.25	7.284601	1.5625	−3.37375
1907–08	−2.167	+1.14	4.695889	1.2996	−2.47038
1908–09	+2.135	−1.50	4.558225	2.2500	−3.20250
1909–10	−3.237	+3.79	10.478169	14.3641	−12.26823
1910–11	+1.604	+.60	2.572816	.3600	+.96240
1911–12	+4.084	−4.79	16.679056	22.9441	−19.56236
1912–13	−1.990	+1.44	3.960100	2.0736	−2.86560
1913–14	+.453	+1.62	.205209	2.6244	+.73386
1914–15	+1.979	−5.20	3.916441	27.0400	−10.29080
1915–16	−4.943	+1.73	24.433249	2.9929	−8.55139
1916–17	+.258	+2.18	.066564	4.7524	+.56244
1917–18	−.148	+3.03	.021904	9.1809	−.44844
1918–19	+.739	−.34	.546121	.1156	−.25126
1919–20	−.620	+2.27	.384400	5.1529	−1.40740
1920–21	+2.019	−6.02	4.076361	36.2404	−12.15438
1921–22	−5.486	+3.63	30.096196	13.1769	−19.91418
1922–23	+1.808	+2.86	3.268864	8.1796	+5.17088
	−22.847	−25.55	140.548313	208.6688	−117.37216
	+22.486	+31.79			
	−.361	+6.24			

TABLE 108 — *Continued*

$$c_x = \frac{-.361}{22} = -.0164$$

$$c_x^2 = .00026896$$

$$c_y = \frac{+6.24}{22} = +.284$$

$$c_y^2 = .080656$$

$$\sigma_x = \sqrt{\frac{140.548313}{22} - .00027}$$

$$= \sqrt{6.38829}$$

$$\sigma_x = 2.528$$

$$\sigma_y = \sqrt{\frac{208.6688}{22} - .08066}$$

$$\sigma_y = \sqrt{9.40429}$$

$$\sigma_y = 3.067$$

$$p = \frac{\Sigma XY}{N} - c_x c_y$$

$$= \frac{-117.37216}{22} - (.284 \times -.0164)$$

$$= -5.3304404$$

$$r = \frac{p}{\sigma_x \sigma_y}$$

$$r = \frac{-5.3304404}{7.753376}$$

$$= -.687$$

There are several ways in which the problem of comparing short term fluctuations may be attacked. The absolute differences between successive items in two series may be correlated, or these differences may be expressed as percentages or ratios. Table 108 illustrates the procedure employed in measuring the correlation between the absolute fluctuations from year to year (first differences) of cotton production and cotton prices. The original values from which the items in columns (2) and (3) are derived are given in Table 103.

The process of computing r is identical with that employed in preceding examples, when deviations were measured from an arbitrary origin. The arbitrary origin in this case is zero, but corrections must be made in the various values since the algebraic sum of the given figures is not zero in either case. A value of $-.687$ is secured for r.

The equation of regression and the value of S_y as computed from the usual formulas are

$$y = .8335x$$

$$S_y = 2.23 \text{ cents}$$

A comparison of the different results secured in the preceding examples relating to cotton throws some inter-

esting light upon the general problem of correlation. In fact we have measured the correlation between three different things which are not strictly comparable — deviations from third degree parabolas, deviations from straight lines of trend, and year to year fluctuations in the production and price of cotton. Yet, if we were seeking to estimate the price of cotton which would accompany a given crop, an estimate might be based upon any one of the three studies, the results of which are given below.

	r	S_y
Correlation of cycles in cotton production and prices (deviations measured from third degree parabolas)	— .567	1.66 cents
Correlation of cycles, same data (deviations measured from straight lines of trend)	— .608	1.72 cents
Correlation of year to year fluctuations, same data	— .687	2.23 cents

The value of r in the latest example is greater than the values secured in the other cases, though the standard error is also larger. The reason for this apparent contradiction has been suggested above; the standard deviation of the year-to-year fluctuations is greater than the standard deviations about the linear and parabolic trends.

At first glance it would appear that the error of estimate would be least if based upon the results secured when deviations from third degree parabolas were correlated, and greatest when based upon the study of year-to-year movements. But there is a concealed assumption in the first case, the assumption that the lines of trend of both prices and production may be projected beyond the period studied. There is an immeasurable margin of error in this assumption, and the standard error of estimate, accordingly, does not give a true measure of the probabilities involved. The same is true with regard to the linear trends. No such assumption is involved in the measure based upon year-to-year fluctuations, though a single exceptional year might introduce greater difficulties here than in either of the other cases.

REFERENCES

MOORE, H. L. *Economic Cycles: Their Law and Cause.*

MOORE, H. L. *Forecasting the Yield and the Price of Cotton.*

MOORE, H. L. *Generating Economic Cycles.*

PERSONS, W. M. *Correlation of Time Series.* Journal of the American Statistical Association, June, 1923. (This article is also published in Rietz, H. L., *Handbook of Mathematical Statistics* 150–165.)

PERSONS, W. M. *Indices of Business Conditions.* Review of Economic Statistics, Prel. Vol. I. 1919.

PERSONS, W. M. *The Variate Difference Correlation Method and Curve Fitting.* Quarterly Publications of the American Statistical Association, June, 1917.

SNOW, E. C. *Trade Forecasting and Prices* (with discussion). Journal of the Royal Statistical Society, May, 1923 (332–398).

STAMP, J. C. *The Effect of Trade Fluctuations Upon Profits* (with discussion). Journal of the Royal Statistical Society, July, 1918 (563–608).

YULE, G. U. *On the Time Correlation Problem, with Especial Reference to the Variate Difference Correlation Method* (with discussion). Journal of the Royal Statistical Society, July, 1921 (497–537).

CHAPTER XII

THE MEASUREMENT OF RELATIONSHIP: NON-LINEAR CORRELATION

In the preceding chapters the discussion has been confined to cases in which the relationship between two variables may be described by a straight line. The coefficient of correlation, r, is a measure of the degree to which two variables approach a linear relationship and it is significant only when a straight line gives a good fit to the points representing the paired values of X and Y.

In fitting curves to time series, as explained in an earlier section, it is found that in many cases the trend is non-linear, and that a curve of higher degree is needed. The same thing is true in the field of our present discussion. It is possible to have a high degree of correlation between two variables when a straight line does not describe the relationship. In such a case there would be considerable scatter about the straight line of best fit, and the value of r would be misleadingly low. If a curve representing the real relationship could be fitted, the scatter would be materially reduced and the true correlation could be measured. The figures presented in Table 109 on page 433 illustrate such a case.

These data are plotted in Fig. 79.

Two different curves have been fitted to the points plotted in this figure. One is a straight line having the equation

$$Y = 5.038 + .0886X$$

in which Y represents yield, in tons per acre, and X represents depth of irrigation water applied, in inches. The degree of relationship between the two variables, as de-

TABLE 109

Alfalfa Yield and Irrigation

(Summary of Investigations at Davis, California, 1910–1915) [1]

Depth of Water applied (inches)	*Yield in tons per acre*						*Averages*
	1910	1911	1912	1913	1914	1915	
0	3.85	5.94	5.52	2.75	2.89	2.35	3.88
12	4.78	7.52	6.51	4.31	5.83	4.84	5.63
18	*	*	7.02	5.69	8.02	6.46	6.80
24	6.00	8.38	8.32	6.89	9.96	7.96	7.92
30	7.53	9.54	9.43	7.97	11.06	8.32	8.98
36	7.58	9.33	9.38	8.22	12.48	8.63	9.27
48	8.45	9.52	8.63	8.83	10.62	8.05	9.02
60	*	*	10.17	7.25	10.70	5.55	8.42

* *No observation*

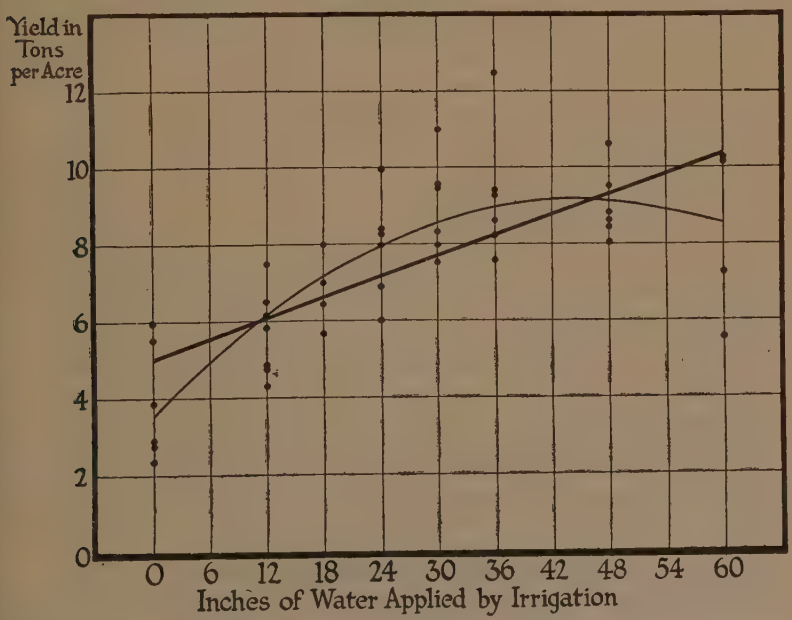


FIG. 79. — Scatter Diagram Showing the Relation between Alfalfa Yield and Irrigation Water Applied, with Two Lines of Regression

[1] This table is taken from "The Economical Irrigation of Alfalfa in the Sacramento Valley" by S. H. Beckett and R. D. Robertson, *Bull. No. 280*, Agricultural Experiment Station, Univ. of California, May, 1917.

scribed by this line, is indicated by the coefficient of correlation, r, which has a value of $+.68$.

An inspection of the figure shows clearly that the straight line does not give the best possible fit. It is certain, therefore, that r does not furnish a valid measure of the degree of relationship between alfalfa yield and depth of irrigation water.

Parabolic Relationship

The other curve is a second degree parabola, fitted by the method of least squares. The equation to this curve is

$$Y = 3.55 + .252X - .002816X^2$$

It is obvious that the effect of increasing irrigation upon alfalfa yield is described much more accurately by this latter curve than by the straight line. The most important result of these investigations was the determination of the point at which alfalfa yield began to fall off with increased applications of water, and the straight line fails to indicate any such decline.

As the equation of relationship, therefore, we should use the parabolic rather than the linear form. The standard error, S_y, which is a necessary accompanying measure, may be calculated by measuring the deviation of each value from the corresponding computed value, and determining the root-mean-square of these deviations. This procedure is illustrated in the following table. The figures for normal yield which are given in this table are computed from the parabolic equation given above.

TABLE 110

Comparison of Actual and Computed Alfalfa Yield

(1)	(2)	(3)	(4)	(5)
Depth of Irrigation Water	*Actual Yield*	*Normal Yield, as computed from parabolic equation*	*Deviation of actual from normal* (2)–(3)	
X	Y	Y_c	d	d^2
0	3.85	3.55	+.30	.0900
0	5.94	3.55	+2.39	5.7121
0	5.52	3.55	+1.97	3.8809
0	2.75	3.55	−.80	.6400
0	2.89	3.55	−.66	.4356
0	2.35	3.55	−1.20	1.4400
12	4.78	6.16	−1.38	1.9044
12	7.52	6.16	+1.36	1.8496
12	6.51	6.16	+.35	.1225
12	4.31	6.16	−1.85	3.4225
12	5.83	6.16	−.33	.1089
12	4.84	6.16	−1.32	1.7424
18	7.02	7.17	−.15	.0225
18	5.69	7.17	−1.48	2.1904
18	8.02	7.17	+.85	.7225
18	6.46	7.17	−.71	.5041
24	6.00	7.97	−1.97	3.8809
24	8.38	7.97	+.41	.1681
24	8.32	7.97	+.35	.1225
24	6.89	7.97	−1.08	1.1664
24	9.96	7.97	+1.99	3.9601
24	7.96	7.97	−.01	.0001
30	7.53	8.57	−1.04	1.0816
30	9.54	8.57	+.97	.9409
30	9.43	8.57	+.86	.7396
30	7.97	8.57	−.60	.3600
30	11.06	8.57	+2.49	6.2001
30	8.32	8.57	−.25	.0625
36	7.58	8.97	−1.39	1.9321
36	9.33	8.97	+.36	.1296
36	9.38	8.97	+.41	.1681
36	8.22	8.97	−.75	.5625
36	12.48	8.97	+3.51	12.3201
36	8.63	8.97	−.34	.1156
48	8.45	9.15	−.70	.4900
48	9.52	9.15	+.37	.1369

(*Continued on page 436*)

TABLE 110 (*Continued*)

Comparison of Actual and Computed Alfalfa Yield

(1)	(2)	(3)	(4)	(5)
48	8.63	9.15	−.52	.2704
48	8.83	9.15	−.32	.1024
48	10.62	9.15	+1.47	2.1609
48	8.05	9.15	−1.10	1.2100
60	10.17	8.53	+1.64	2.6896
60	7.25	8.53	−1.28	1.6384
60	10.70	8.53	+2.17	4.7089
60	5.55	8.53	−2.98	8.8804
			+24.22	80.9871
			−24.21	

Inserting the sum of the squared deviations in the formula

$$S_y = \sqrt{\frac{\Sigma d^2}{N}}$$

we have

$$S_y = \sqrt{\frac{80.9871}{44}} = 1.36$$

THE INDEX OF CORRELATION

We need now the third value, the abstract measure of degree of relationship. In dealing with cases of linear relationship in the preceding chapter we found that such a measure, the coefficient of correlation, could be derived from known values of S_y and σ_y. An analogous measure may be derived in the same way in cases of non-linear relationship, such as that found in the present problem. Since the term *coefficient of correlation* and the symbol r refer only to cases of linear regression, we may term this general measure the *index of correlation*, and use the symbol ρ (Rho) to represent it.[1]

[1] This symbol has been employed by Spearman to represent the coefficient of correlation when based upon the squares of differences in rank. This usage has no apparent significance in ordinary economic analysis, however, and confusion is not likely to arise from the employment of the symbol for the present purpose.

As a general formula for the index of correlation we have[1]

$$\rho^2_{yx} = 1 - \frac{S_y^2}{\sigma_y^2}$$

The value of S_y has been derived above. The value of σ_y, computed by familiar methods, is found to be 2.27. Substituting in the formula for ρ we have

$$\rho_{yx} = \sqrt{1 - \frac{1.8406}{5.177}}$$

$$= .80$$

This value is materially greater than that of the coefficient of correlation for the same data. The value of r is $+.68$. The difference is due to the fact that the second degree parabola constitutes a much better fit to the data than the straight line. The correlation is distinctly non-linear, and r is an inappropriate measure of correlation.

The Meaning of the Index of Correlation

It is important that the significance and the limitations of ρ be understood. Its value depends upon the relation between the scatter about the fitted line and the scatter about the arithmetic mean of the Y's. In the case of a straight line, ρ and r are identical, r being a special case of ρ. The limits of ρ are 0 and 1, a value of 0 indicating that if there is a relationship between the two variables it cannot be described by the particular equation employed. A value of 1 indicates that the relationship, as described by the equation employed, is a perfect one. For curves

[1] With X dependent this formula becomes

$$\rho^2_{xy} = 1 - \frac{S_x^2}{\sigma_x^2}$$

The first of the two subscripts refers always to the dependent variable, the second to the independent. It is essential that these be shown, for ρ would not necessarily be the same with X dependent as with Y dependent. Such a distinction is not necessary in the case of linear correlation, for r is the same no matter which variable be dependent.

of higher degree no positive or negative sign should be attached to ρ, for the relationship might be positive over part of the range and negative over other parts, as in the alfalfa example given above.

The index of correlation, ρ, has no significance unless the type of curve to which it applies be named in each case. The meaning of r in this respect is always clear, for it is understood that it relates always to a straight line, but confusion would arise in the case of ρ unless the type of curve were specifically mentioned. The index of correlation may be looked upon as a measure of the adequacy of a curve of a given type to describe the relationship between two variables.

It is, of course, always possible to secure a curve which will pass through any number of points if the constants in the equation be equal to the number of points. In such a case ρ would, of necessity, be equal to 1, but this value would have no significance. In any employment of mathematical functions there is this limit of absurdity, when the number of constants is equal to the number of points, and ρ would merely reflect this absurdity. The ordinary principles of curve fitting must be kept in mind in using such an index as this. It must never be taken to have an absolute significance, standing by itself. Its significance is always relative, referring to the particular function employed. This fact, which is true of every measure of correlation, is frequently overlooked, and invalid and fallacious conclusions reached as a result.

A Short Method of Computing the Index of Correlation

The standard error and the index of correlation were computed by a rather laborious method in the above example, in order that there might be no misunderstanding of their precise meaning. The burden of calculation may

be materially reduced, however, by taking advantage of the relationships which were disclosed in dealing with r. For a curve of the potential series

$$Y = a + bX + cX^2 + dX^3 \ldots$$

the formula for S_y is derived by a simple extension of that employed in the case of the straight line. As a general formula for a series of this type, we have

$$S_y^2 = \frac{\Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY) - c\Sigma(X^2Y) - d\Sigma(X^3Y) - \ldots}{N}$$

Similarly, the formula for r may be extended to give a general formula for ρ applicable to any equation of this general type. This formula[1] is

$$\rho^2_{yx} = \frac{a\Sigma(Y) + b\Sigma(XY) + c\Sigma(X^2Y) + d\Sigma(X^3Y) + \ldots - Nc_y^2}{\Sigma(Y^2) - Nc_y^2}$$

In the special case in which the origin is at the mean of the Y's, $\Sigma(y) = 0$ and $c_y = 0$, and the formula reduces to

$$\rho^2_{yx} = \frac{b\Sigma(Xy) + c\Sigma(X^2y) + d\Sigma(X^3y) + \ldots}{\Sigma(y^2)}$$

The characteristics of the formulas for S and ρ should be noted. The only values required in securing these measures are the constants in the equation which describes the average relationship, certain values which have been used in the process of fitting and, in addition, $\Sigma(Y^2)$ and c_y^2. Thus, as direct by-products of the fitting process, we have the values of S and ρ, the two measures which are needed to supplement the regression equation in securing a complete description of the relationship between the two variables in question. The equation describes the average relationship. The standard error, S, is a measure of the reliability of estimates based upon this equation, and ρ is an abstract index of the degree of relationship, in so far as

[1] See Appendix A for a discussion of the derivation of this formula.

that relationship can be described by the particular curve employed.

The application of these formulas may be illustrated with reference to the problem of alfalfa yield. The following values, derived from the data of Table 109 and from the fitting process, are required for this purpose:

$$a = 3.5468 \qquad \Sigma(X^2Y) = 407{,}448.00$$
$$b = .2520 \qquad c_y^2 = 55.9504$$
$$c = -.0028162$$
$$\Sigma(Y) = 329.03 \qquad \Sigma(Y^2) = 2688.3129$$
$$\Sigma(XY) = 10{,}269.96 \qquad N = 44$$

Substituting in the formula for the standard error for a second degree parabola,

$$S_y^2 = \frac{\Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY) - c\Sigma(X^2Y)}{N}$$

we have

$$S^2_y = \frac{2688.3129 - (3.5468 \times 329.03) - (.2520 \times 10{,}269.96) - (-.0028162 \times 407{,}448)}{44}$$

$$= \frac{80.7345}{44}$$

$$= 1.8349$$

$$S_y = 1.36$$

The index of correlation, for a curve of this type, is computed from the equation

$$\rho^2_{yx} = \frac{a\Sigma(Y) + b\Sigma(XY) + c\Sigma(X^2Y) - Nc_y^2}{\Sigma(Y)^2 - Nc_y^2}$$

Substituting the appropriate values, we have

$$\rho^2_{yx} = \frac{145.7608}{2688.3129 - (44 \times 55.9504)}$$

and

$$\rho_{yx} = .80$$

These methods of deriving S and ρ are applicable over a wide field by a simple adaptation of the formulas to the

particular equations that may be employed in given instances. Further illustrations are given in the succeeding chapter, while this general method is explained in more detail in Appendix A.

The Correlation Ratio

A third distinctive measure of correlation remains to be described. This is the *correlation ratio*, devised by Karl Pearson and represented by the symbol η (Eta). This measure may be looked upon as a special case of ρ, but somewhat different methods are employed in its computation.

We have seen that in all cases the degree of relationship between two variables, as described by a curve of a given type, may be determined from the formula

$$\text{Measure of correlation} = \sqrt{1 - \frac{S_y^2}{\sigma_y^2}}$$

The coefficient of correlation, r, is just such a measure, when S_y represents the standard deviation about a straight line. The index of correlation, ρ, is a general measure of the same type. The *correlation ratio* is precisely the same sort of measure, S_y in this case representing the standard deviation about a line passing through the mean of every column in the correlation table. We have, in effect, increased the number of constants in the equation of the curve to be fitted until the number is equal to the number of columns. If the mid-points of all the columns lie on a straight line, the correlation ratio and the coefficient of correlation will be equal. If the mid-points of the columns do not lie on a straight line, the correlation ratio will be greater than the coefficient of correlation.

No new principle is involved, therefore, in the concept of the correlation ratio. It is employed when the regression is non-linear. It measures the degree of relationship between two variables, in so far as this relationship may be

described by a curve passing through the mean of every column. If the relationship is perfect, if there is no scatter about the curve fitted in this way, η will have a value of 1. If there is no relationship, if the scatter about the curve is as great as the dispersion about the mean of the Y's, η will have a value of zero.

The formula generally employed in the computation of the correlation ratio differs somewhat from that given above. To represent the standard deviation about the line joining the means of the columns, the symbol σ_{ay} is employed, instead of S_y. Its meaning is precisely the same as that of S_y, as employed above, except that σ_{ay} refers always to a correlation table.

The formula may be written

$$\eta_{yx} = \sqrt{1 - \frac{\sigma_{ay}^2}{\sigma_y^2}}$$

When *eta* is written as above (η_{yx}) it refers to the regression of Y on X (Y dependent). When it is written η_{xy} it refers to the regression of X on Y (X dependent), and its value depends upon the scatter about a line joining the means of the rows. Unlike r, which has the same value for both regressions, η_{yx} and η_{xy} will have different values unless the regression be linear.

The Computation of the Correlation Ratio

The following correlation table shows the general relation between the amount of nitrogen, in pounds per acre, used as fertilizer in certain agricultural experiments, and the corresponding yield of wheat, in bushels per acre.[1] The points are plotted in Fig. 80 on page 444.

[1] This table is based upon experiments described by E. Davenport ("Comparative Agriculture" in Bailey's *Cyclopedia of American Agriculture*). The actual figures used have been arbitrarily chosen for the purpose of the present illustration, but Davenport's experiments have demonstrated the existence of a law similar to the one here assumed.

TABLE 111

Correlation Table showing the Relation between Wheat Yield per Acre and amount of Nitrogen used as Fertilizer

		X — Nitrogen applied in pounds per acre										
		0–19.9	20–39.9	40–59.9	60–79.9	80–99.9	100–119.9	120–139.9	140–159.9	160–179.9	Totals	Means of Rows
Y — Wheat yield in bushels per acre	32–35.9				5	16	12	4	5	2	44	107.27
	28–31.9			1	20	21	8	4	1		55	88.91
	24–27.9			16	19						35	60.86
	20–23.9			13							13	50.0
	16–19.9		12								12	30.0
	12–15.9		8								8	30.0
	8–11.9	3	5								8	22.50
	4–7.9	10									10	10.0
	0–3.9	8									8	10.0
	Totals	21	25	30	44	37	20	8	6	2	193	
	Means of Columns	5.05	15.12	24.4	28.73	31.73	32.4	32.0	33.33	34.0		

For the computation of η_{yx} by the formula given above we need the values of σ_y and σ_{ay}, the latter being the root-mean-square deviation about the line joining the means of the various columns. The former value may be

obtained readily by methods already familiar. It is possible to compute the quantity σ_{ay} by the method first employed in calculating S_y, that is, by measuring and squaring the deviations of the individual points from the line of regression. In the present case, however, the line describing the relationship passes through the mean of each column, hence these means may be used in place of the "normal"

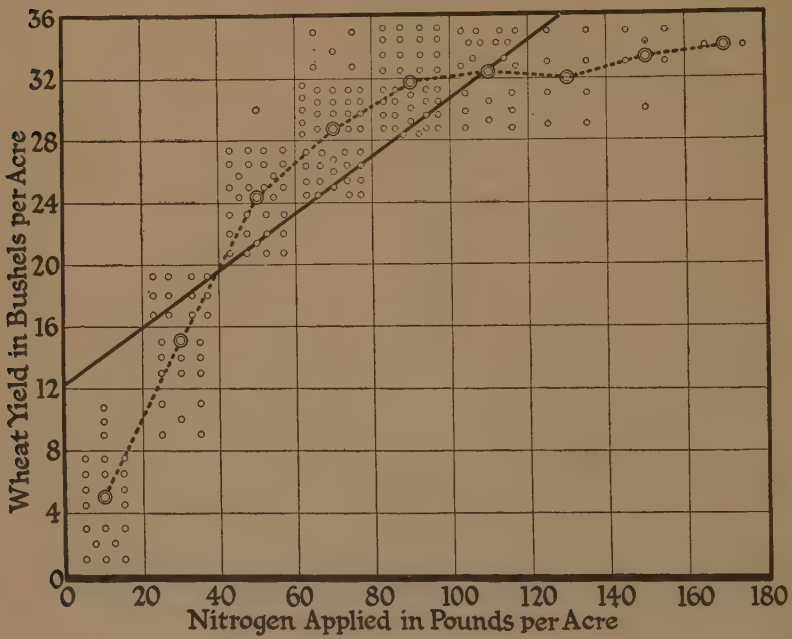


FIG. 80. — Scatter Diagram Showing the Relation between Wheat Yield and Nitrogen applied as Fertilizer, with Straight Line of Regression and Line joining the Means of the Columns

values as computed from an equation of regression. In computing σ_{ay}, therefore, the deviations of the individual items from the means of the various columns are squared, added, the mean determined and the square root extracted, just as in the computation of the standard deviation. Part of the procedure may be illustrated by using the

data in the first column of Table 111. This column contains all items having X-values between 0 and 20. The mean Y-value of the 21 items falling in this column is 5.05; deviations are measured from this value.

TABLE 112

Computation of the Squares of the Deviations about the Mean of an Array

Class-Interval (wheat yield in bu. per acre)			*Deviation from mean of column (5.05)*		
	m	f	d	d^2	fd^2
8 – 11.9	10	3	4.95	24.5025	73.5075
4 – 7.9	6	10	.95	.9025	9.0250
0 – 3.9	2	8	−3.05	9.3025	74.4200
Total					156.9525

The sum of the squared deviations is obtained for each of the other columns in a similar fashion. The standard deviation about the means of all the columns, σ_{ay}, is found to have a value of 2.420. The value of σ_y is 9.188.

Substituting the given values in the formula

$$\eta^2_{yx} = 1 - \frac{\sigma_{ay}^2}{\sigma_y^2}$$

$$\eta^2_{yx} = 1 - \frac{(2.42)^2}{(9.188)^2}$$

$$= 1 - .0694$$

$$= .9306$$

$$\eta_{yx} = .965$$

This is the value of the correlation ratio, measuring the degree of scatter about a line running through the means of the columns. Its significance is discussed below.

The method of calculation employed in the preceding example may be materially shortened. Let σ_{my} represent the standard deviation of the means of the various columns about the arithmetic mean of all the Y's. In computing

this value the mean of each column is weighted by the number of items in that column. It may be shown[1] that

$$\sigma_{ay}^2 = \sigma_y^2 - \sigma_{my}^2$$

Substituting for σ_{ay}^2 in the equation

$$\eta^2_{yx} = 1 - \frac{\sigma_{ay}^2}{\sigma_y^2},$$

we secure

$$\eta^2_{yx} = 1 - \left(\frac{\sigma_y^2 - \sigma_{my}^2}{\sigma_y^2}\right)$$

$$= \frac{\sigma_{my}^2}{\sigma_y^2}$$

$$\eta_{yx} = \frac{\sigma_{my}}{\sigma_y}$$

Since σ_{my} may be much more easily determined than σ_{ay} the value of η is generally computed from this formula. The data of Table 111 may be used to exemplify the process.

[1] The following proof is adapted from Yule.

Given a series with mean M made up of two component series with means M_1 and M_2. N, the total number of observations, is equal to $N_1 + N_2$, the sum of the observations in the two component series. What is the relation between σ, σ_1 and σ_2? If we let

$$M_1 - M = c_1$$

then for S_1^2, the mean-square deviation of the observations in the first of the two component series, measured from M as origin, we have

$$S_1^2 = \sigma_1^2 + c_1^2$$

Similarly

$$S_2^2 = \sigma_2^2 + c_2^2$$

But $N_1S_1^2$ is equal to the sum of the squares of the deviations, about M, of the items in the first of the component series, and $N_2S_2^2$ is equal to the sum of the squares of the deviations, about M, of the items in the second of the two component series. Therefore

$$\sigma^2 = \frac{N_1S_1^2 + N_2S_2^2}{N}$$

and
$$N\sigma^2 = N_1S_1^2 + N_2S_2^2 \qquad (1)$$

But
$$S_1^2 = \sigma_1^2 + c_1^2 \text{ and } S_2^2 = \sigma_2^2 + c_2^2$$

therefore
$$N\sigma^2 = N_1(\sigma_1^2 + c_1^2) + N_2(\sigma_2^2 + c_2^2) \qquad (2)$$

In the present case we have the major series with mean represented by M_y, and a number of component series (the items arranged by columns) with means represented by m_{y1}, etc. Let S_{ay} represent the standard deviation of any column

TABLE 113

Illustrating the Computation of the Correlation Ratio

Type of Array (X-value of items in column) (pounds)	*Mean value of Y-items in column (bushels)*	*Deviation from Mean of all Y's (25.005)*	*Square of Deviation*	*Frequency*	
	m_y	d	d^2	f	fd^2
10	5.05	−19.955	398.202	21	8,362.242
30	15.12	−9.885	97.713	25	2,442.825
50	24.40	−.605	.366	30	10.980
70	28.73	+3.725	13.876	44	610.544
90	31.73	+6.725	45.226	37	1,673.362
110	32.40	+7.395	54.686	20	1,093.720
130	32.00	+6.995	48.930	8	391.440
150	33.33	+8.325	69.306	6	415.836
170	34.00	+8.995	80.910	2	161.820
Total				193	15,162.769

$$\sigma_{my} = \sqrt{\frac{15{,}162.769}{193}}$$

$$= 8.864$$

---

of Y's about the mean of that column. Then we have a number of component series, with standard deviations S_{ay1}, etc., and with means differing from the mean of all the Y's by $M_y - m_{y1}$, etc. Substituting in equation (2) derived above, we have

$$N\sigma_y^2 = n_1[S_{ay1}^2 + (M_y - m_{y1})^2] + n_2[S_{ay2}^2 + (M_y - m_{y2})^2] + \dots \quad (3)$$

$$N\sigma_y^2 = \Sigma n[S_{ay}^2 + (M_y - m_y)^2] \quad (4)$$

But $$N\sigma_{ay}^2 = \Sigma(n \cdot S_{ay}^2)$$

for, in each column, $$S_{ay}^2 = \frac{\Sigma d^2}{n}$$

since d represents a deviation from the mean of that column. Then, for all columns,

$$\sigma_{ay}^2 = \frac{\Sigma d^2}{N} = \frac{\Sigma(n \cdot S_{ay}^2)}{N}$$

Substituting in equation (4)

$$N\sigma_y^2 = N\sigma_{ay}^2 + \Sigma n(M_y - m_y)^2 \quad (5)$$

By definition of the standard deviation of the means of the columns

$$\sigma_{my}^2 = \frac{\Sigma n(M_y - m_y)^2}{N}$$

Therefore, from (5), $$\sigma_y^2 = \sigma_{ay}^2 + \sigma_{my}^2 \quad (6)$$

Substituting the given values in the formula

$$\eta_{yx} = \frac{\sigma_{my}}{\sigma_y}$$

we have

$$\eta_{yx} = \frac{8.864}{9.188}$$
$$= .965$$

The process of computing the correlation ratio may be briefly summarized:

1. Arrange the items in the form of a correlation table.
2. Find the arithmetic mean of all the Y-items in each column. (i.e. Find the arithmetic mean of each Y-array of type X.)
3. Compute the arithmetic mean of all the Y's.
4. Measure the deviation of the mean of each column from the mean of all the Y's. Square each of these deviations and multiply by the number of items in the given column. Get the sum of the squared deviations.
5. Divide this sum by the total number of items and extract the square root of the result. This gives the value of σ_{my}.
6. Compute σ_y.
7. Divide σ_{my} by σ_y. The quotient is η_{yx}.

The value of the correlation ratio of X on Y may be similarly computed, substituting the proper values in the formula

$$\eta_{xy} = \frac{\sigma_{mx}}{\sigma_x}$$

The symbol σ_{mx} represents the standard deviation of the means of the various *rows* about the mean of all the X's. The value of the correlation ratio of X on Y depends upon the amount of scatter (horizontally) about the line joining the means of the rows. Its value will generally be different from that of the correlation ratio of Y on X. In the present case the value of η_{xy} is found to be .824. As the line of relationship approaches the linear form the two correlation ratios approach identity.

Like r, η can never exceed 1, this value being secured when there is no dispersion about the line joining the means of the columns (or rows). From the formula

$$\eta_{yx} = \frac{\sigma_{my}}{\sigma_y}$$

it is evident that the value of the correlation ratio is zero when σ_{my} is zero. This is the case when the mean of each column has the same value as the mean of all the Y's. Such a condition is found when an increase or decrease in the value of the X-variable brings no corresponding change in the value of the Y-variable. This means that in each column of the correlation table there is a distribution of cases similar to the general distribution of Y's. When this is true there is clearly no relation between the two variables.

The correlation ratio, it should be noted, never has a negative value. It is possible to determine by inspection of the correlation table, however, whether the relation between two variables is direct, or inverse, or a varying one.

The coefficient of correlation has one distinct advantage, as compared with the correlation ratio, in that when its value and the values of the two standard deviations are known the equations to the lines of regression may be readily determined. This is not true of η. To get a quantitative expression for the "law" of relationship between two variables, when η has been computed, an additional calculation for the purpose of fitting a curve to the means of the arrays would be necessary.

Correction of the Correlation Ratio

The use of η is only possible when the data are numerous, and can be arranged in the form of a correlation table. If a limited number of items should be so arranged, and it chanced that there was but one item in each column, the two measures σ_{my} and σ_y would be identical and η would

necessarily have a value of 1. Computed from a very small number of cases and employing a large number of classes, the correlation ratio would be meaningless.

Karl Pearson has suggested a correction for the raw correlation ratio to offset, in particular, errors due to the employment of too fine a grouping (i.e. too many classes) for the data. If we represent by κ the number of arrays, the desired value is secured from the formula

$$\text{Corrected } \eta^2 = \frac{\eta^2 - \frac{(\kappa - 3)}{N}}{1 - \frac{(\kappa - 3)}{N}}$$

Applying this correction to the results secured in the preceding example, we have

$$\text{Corrected } \eta^2 = \frac{(.965)^2 - \frac{(9 - 3)}{193}}{1 - \frac{(9 - 3)}{193}}$$

$$= .929$$

$$\text{Corrected } \eta = .964$$

The correction is very slight in the present case, but if N were small or κ very large it would reduce the given value materially.

Relation between the Correlation Ratio and the Coefficient of Correlation

When the relation between two variables is absolutely linear the line running through the means of the columns corresponds, of course, to the line upon which the coefficient of correlation is based. When this is the case η and r have the same value. As the relationship between the two variables departs from the linear form the values secured for η and r differ, η being always greater than r. This results from the fact that the scatter about a line joining the means of the columns will always be less than

the scatter about a straight line fitted to these points, except when the straight line passes through every mean point. And the less the scatter about the line expressing the average relationship the greater the value of the measure of correlation. Thus for the alfalfa problem it was found that r has a value of + .68, and that an index of correlation based upon a second degree parabola has a value of .80. The correlation ratio for the same material is .83. For the data of Table 111 the value of η_{yx} was found to be .964; the value of r is + .793, the difference between the two being marked. The reason for the difference is found in Fig. 80, in which the straight line of regression of Y on X and the line joining the means of the columns are shown. The regression departs materially from linearity, and the scatter about the straight line of regression is much greater than the scatter about the line joining the means.

The relation between r and η affords a convenient test of linearity in a given instance, since the two values will be identical when the regression is strictly linear, and will differ the more widely the greater the departure from the linear form. The general test for linearity is

$$\zeta = \eta^2 - r^2$$

Even in a case of linear regression it is probable that η and r will differ somewhat because of fluctuations due to chance alone. A material difference, as reflected in the magnitude ζ (Zeta), indicates that a straight line does not describe the relationship in question and that r is not a suitable measure of correlation.

In the example given above, in which η equals .964 and r equals .793, the measure ζ has a value of .300. This is large enough to indicate conclusively that the regression is non-linear. In a later section[1] the method of determining the significance of a given value of ζ is explained.

[1] Chapter XVI.

REFERENCES

Bowley, A. L. *Elements of Statistics* (365–367).

Kelley, Truman L. *Statistical Method* (238–245).

Pearl, Raymond. *Medical Biometry and Statistics* (311–318).

Pearson, Karl. *Mathematical Contributions to the Theory of Evolution (XIV). On the General Theory of Skew Correlation and Non-Linear Regression.* Draper's Company Research Memoirs, Biometric Series II. 1905.

Notes on the History of Correlation. Biometrika, Vol. 13, 1920 (25–45).

On a Correction Needful in the Case of the Correlation Ratio. Biometrika, Vol. 8, 1911 (254–256).

On the Correction Necessary for the Correlation Ratio. Biometrika, Vol. 14, 1923 (412–417).

Rietz, H. L. (editor) *Handbook of Mathematical Statistics* (129–131).

Yule, G. U. *An Introduction to the Theory of Statistics* (204–207).

CHAPTER XIII

THE MEASUREMENT OF RELATIONSHIP AND THE PROBLEM OF ESTIMATION

It is no great exaggeration to say that quantitative method in economics and business centers about the problem of estimation. Equations of regression, measures of standard error and coefficients of correlation are of interest largely because of their bearing upon the practical problems of determining probable production, probable price, probable business changes. It should not be understood from this that the problem of estimation relates only to attempts to forecast future changes. We make an estimate whenever we seek to determine the most probable value from a number of different observations, or whenever we employ an equation which describes the relation between two or more variables. The value of statistical technique rests in large part upon its practical utility in the making of estimates.

This object has been definitely to the fore in the preceding chapters, which dealt with methods by which the value of one variable might be estimated from a given value of another. We may, at this point, briefly summarize certain assumptions upon which the validity of this method rests.

Some Assumptions Involved in the Making of Estimates

In earlier chapters it has been pointed out that the most probable value of a series of observations is their arithmetic mean. Given a normal distribution about the mean, the standard deviation affords an exact measure of the probabilities involved in basing estimates upon the mean.

Similarly, the standard error of estimate affords an exact measure of the probabilities involved in basing estimates upon an equation of regression, again upon the assumption that the distribution about the line of regression is normal. The significance and usefulness of the equation of regression may be determined by comparing the standard error of estimate of a given variable with the standard deviation.

From the relation between these two values, moreover, an abstract measure of relationship, the coefficient or index of correlation, may be computed. This coefficient, or index, is a thoroughly valid and accurate measure only if the distribution about the line of regression and the distribution about the mean are normal, or approximately so. Pronounced departures from the normal type lessen the significance of these measures.

In the foregoing discussion we have been concerned with *arithmetic* values throughout. In speaking of estimates based upon the mean we referred to the arithmetic mean. The distributions about the mean and about the line of regression are assumed to be normal when deviations are measured arithmetically. The standard deviation and the standard error of estimate are in arithmetic terms, referring to absolute values. But may we assume that all the distributions we deal with in economic analysis are of the arithmetic type? Should estimates be made and errors of estimate measured only in arithmetic terms? If they should not be so limited, are the methods developed above capable of adaptation to other distributions? These questions may best be answered in terms of a specific problem.

A Problem of Estimation

In Table 114 the production and price of oats in the United States from 1881 to 1913 are recorded. Appropriate lines of trend were fitted to these series and the ratios of the actual values of the items in each series to the trend values determined.

TABLE 114

Production and Price of Oats in the United States

Year	*Production of oats in U. S. (millions of bu.)*	*Straight line trend of production*[1]	*Ratio of actual production to trend value*	*Price of oats in Chicago (cents per bu.)*	*Straight line trend of price*[2]	*Ratio of actual price to trend value*
1881	416	448	.929	47	36.0	1.30
1882	488	471	1.036	37	35.3	1.05
1883	571	494	1.156	31	34.6	.90
1884	583	517	1.128	29	34.0	.85
1885	629	540	1.165	28	33.2	.84
1886	624	563	1.108	25	32.5	.77
1887	659	586	1.124	30	31.2	.96
1888	701	609	1.151	24	30.5	.79
1889	751	632	1.188	24	29.8	.81
1890	523	655	.798	43	29.0	1.48
1891	738	678	1.088	31	28.3	1.10
1892	661	701	.943	30	27.5	1.09
1893	639	724	.882	31	26.8	1.16
1894	662	747	.886	28	26.1	1.07
1895	824	770	1.070	19	25.3	.75
1896	780	793	.983	18	23.6	.76
1897	791	816	.969	24	25.0	.96
1898	843	839	1.005	25	26.4	.95
1899	926	862	1.074	23	27.8	.83
1900	914	885	1.033	25	29.2	.86
1901	778	908	.857	42	30.6	1.37
1902	1053	931	1.131	33	32.0	1.03
1903	869	954	.911	38	33.4	1.14
1904	1009	977	1.033	30	34.8	.86
1905	1090	1000	1.090	31	36.2	.86
1906	1036	1023	1.013	39	37.6	1.04
1907	805	1046	.770	51	39.0	1.31
1908	851	1069	.796	52	40.4	1.29
1909	1068	1092	.978	43	41.8	1.03
1910	1186	1115	1.064	35	43.2	.81
1911	922	1138	.810	51	44.6	1.14
1912	1418	1161	1.221	37	46.0	.80
1913	1122	1184	.948	41	47.4	.87

[1] This line of trend was fitted to data covering a longer period than that included in the present study.

[2] The entire period has been broken into two parts, 1881 to 1895, and 1896 to 1913. A straight line of trend was fitted by H. B. Killough to the data of each period.

It is desired to measure the relation between these two variables. A hyperbolic curve of the general type $Y = aX^b$ appears to be an appropriate form to employ in describing such a relationship. To fit this curve by the method of least squares, the equation must be reduced to the logarithmic form

$$\log Y = \log a + b \log X$$

The normal equations required in fitting a curve of this type, are

$$\text{I} \quad \Sigma(\log Y) = N \log a + b\,\Sigma(\log X)$$

$$\text{II} \quad \Sigma(\log X \cdot \log Y) = \log a \Sigma(\log X) + b\Sigma(\log^2 X)$$

The values necessary for the solution of these equations are determined from Table 115.[1]

From this table we have

$$N = 33$$

$$\Sigma(\log Y) = -.32849 \qquad \Sigma(\log X \cdot \log Y) = -.1143005$$

$$\Sigma(\log X) = .037535 \qquad \Sigma(\log^2 X) = .096423$$

Substituting in the normal equations, we secure

$$-.32849 = 33 \log a + .037535b$$

$$-.1143005 = .037535 \log a + .096423b$$

Solving

$$\log a = -.00861$$

$$b = -1.18206$$

The required equation is

$$\log Y = (9.99139 - 10) - 1.18206 \log X$$

or

$$Y = .9804X^{-1.18206}$$

This is the equation which describes the average relationship between the production and the price of oats (when the actual figures for each are expressed as ratios to the respective lines of trend). The corresponding curve is plotted in Fig. 82.

[1] I am indebted to Mr. H. B. Killough of the Bureau of Agricultural Economics for permission to use the data presented in Tables 114 and 115. The figures are taken from his comprehensive study of the factors affecting oat prices.

TABLE 115

Computation of Values Required in Fitting a Curve to Data of Oat Production and Prices

EXAMPLE I

(1) Year	(2) Ratio of price to trend Y	(3) Ratio of production to trend X	(4) log Y	(5) log X	(6) $\log^2 Y$	(7) $\log^2 X$	(8) log Y . log X
1881	1.30	.929	.1139434	.9680157−1	.01298310	.001022995	−.0036444
1882	1.05	1.036	.0211893	.0153598	.00044899	.000235923	.0003255
1883	.90	1.156	.9542425−1	.0629578	.00209375	.003963685	−.0028808
1884	.85	1.128	.9294189−1	.0523091	.00498169	.002736242	−.0036920
1885	.84	1.165	.9242793−1	.0663259	.00573362	.004399125	−.0050222
1886	.77	1.108	.8864907−1	.0445398	.01288436	.001983794	−.0050557
1887	.96	1.124	.9822712−1	.0507663	.00031431	.002577217	−.0009000
1888	.79	1.151	.8976271−1	.0610753	.01048021	.003730192	−.0062524
1889	.81	1.188	.9084850−1	.0748164	.00837500	.005597494	−.0068468
1890	1.48	.798	.1702617	.9020029−1	.02898905	.009603432	−.0166852
1891	1.10	1.088	.0413927	.0366289	.00171336	.001341676	.0015162
1892	1.09	.943	.0374265	.9745117−1	.00140074	.000649653	−.0009539
1893	1.16	.882	.0644580	.9454686−1	.00415483	.002973674	−.0035150
1894	1.07	.886	.0293838	.9474337−1	.00086341	.002763216	−.0015446
1895	.75	1.070	.8750613−1	.0293838	.01560968	.000863408	−.0036712
1896	.76	.983	.8808136−1	.9925535−1	.01420540	.000055450	.0008875
1897	.96	.969	.9822712−1	.9863238−1	.00031431	.000187038	.0002425
1898	.95	1.005	.9777236−1	.0021661	.00049624	.000004692	−.0000483
1899	.83	1.074	.9190781−1	.0310043	.00654835	.000961267	−.0025089
1900	.86	1.033	.9344985−1	.0141003	.00429045	.000198818	−.0009236
1901	1.37	.857	.1367206	.9329808−1	.01725316	.004491573	−.0091629
1902	1.03	1.131	.0128372	.0534626	.00016479	.002858250	.0006863
1903	1.14	.911	.0569049	.9595184−1	.00323817	.001638760	−.0023036
1904	.86	1.033	.9344985−1	.0141003	.00429045	.000198818	−.0009236
1905	.86	1.090	.9344985−1	.0374265	.00429045	.001400743	−.0024515
1906	1.04	1.013	.0170333	.0056094	.00029013	.000031465	.0000955
1907	1.31	.770	.1172713	.8864907−1	.01375256	.012884361	−.0133113
1908	1.29	.796	.1105897	.9009131−1	.01223008	.009818214	−.0109580
1909	1.03	.978	.0128372	.9903389−1	.00016479	.000093337	−.0001240
1910	.81	1.064	.9084850−1	.0269416	.00837500	.000725850	−.0024656
1911	1.14	.810	.0569049	.9084850−1	.00323817	.008374812	−.0052076
1912	.80	1.221	.9030900−1	.0867157	.00939155	.007519613	−.0084036
1913	.87	.948	.9395193−1	.9768083−1	.00365792	.000537855	.0014027
Totals	32.83	33.338	17.6715068−18	14.0375350−14	.21721807	.096422642	−.1194567
							+.0051562
							−.1143005

THE STANDARD ERROR OF ESTIMATE IN LOGARITHMIC TERMS

How reliable is this equation? With what degree of confidence may estimates be based upon it? To answer these questions we must compute the standard error, *S*.

Since the fitting process was carried through in terms of logarithms, the standard error may be computed in the same terms. Following the procedure explained in earlier sections with reference to the straight line and the potential series, we may derive the following equation relating to the logarithmic curve just fitted:

$$S^2_{\log y} = \frac{\Sigma(\log^2 Y) - \log a\Sigma(\log Y) - b\Sigma(\log X \cdot \log Y)}{N}$$

Substituting the proper values, we have

$$S^2_{\log y} = \frac{.21721807 - (-.00861 \times -.32849) - (-1.18206 \times -.1143005)}{33}$$

$$= \frac{.07927928}{33}$$

$$S^2_{\log y} = .0024024$$

$$S_{\log y} = .04901$$

The standard error of estimate, in the form of a logarithm, is .04901. As long as we deal with logarithms, this is to be interpreted precisely as is the standard error with respect to other curves. Assuming a normal distribution of logarithms about the curve which describes the average relationship, the chances are 68 out of 100 that the logarithm of a given estimate will not differ from the logarithm of the actual value by more than .04901, 95 out of 100 that the logarithm of the given estimate will not differ from the logarithm of the actual value by more than .09802, and 99.7 out of 100 that the logarithm of the given estimate will not differ from the logarithm of the actual value by more than .14703.

Interpretation of the Standard Error of Estimate; Zones of Estimate

What does this mean in terms of actual values? It means, simply, that we are dealing throughout in terms of ratios instead of absolute figures. The difference between

the logarithms of two numbers is the logarithm of the ratio of one of the original numbers to the other. Thus the absolute value of S in a given case will depend upon the magnitude of the values with which we are dealing. If the user desires to reduce S to absolute values, it must be done always with reference to a given estimate. That is, a given value of X is substituted in the equation of average relationship and the corresponding value of Y estimated. If the logarithmic equation is used, this estimate will be in the form of a logarithm. To the logarithm of the estimate add the value of $S_{\log y}$. The anti-logarithm of the number thus secured will give the upper limit of a zone extending a distance equal to S above the line of regression. From the logarithm of the estimate subtract the value of $S_{\log y}$. The anti-logarithm of the number thus secured will give the lower limit of a zone extending a distance equal to S below the line of regression. The odds are 68 out of 100 that the value of Y in the given case will fall within the limits thus marked out. The absolute limits corresponding to $2S$ and $3S$ may be similarly determined.

The zone thus marked out with respect to a logarithmic curve will differ materially from the similar zones already described in dealing with simple linear equations. In the simple case a zone extending $1S$ on each side of the estimating curve has the same absolute width throughout its length, and is centered always at the line of regression. The logarithmic zone, when measured in natural numbers, is of varying width, and, moreover, is not of the same width on each side of the plotted curve. It is true, however, that the ratios on the two sides of the curve are always equal. That is, the ratio of a value $1S$ less than the computed value to the computed value is the same as the ratio of the latter to a value $1S$ greater. And when the curves are plotted on paper ruled logarithmically, the zone included within a distance $1S$ on each side of the plotted curve takes the symmetrical form found in the earlier and

simpler cases. A person accustomed to thinking in terms of ratios and to the use of logarithmic paper can readily interpret this measure.

The Standard Error of Estimate in Terms of Ratios

Since the ratios are equal throughout, the standard error of estimate may be expressed in ratio terms. In the present example we have

$$S_r = \text{anti-log } S_{\log y} = \text{anti-log } .04901 = 1.12$$

where S_r is used to represent the standard error of estimate in terms of ratios. $S_{\log y}$, as derived above, is positive, hence the ratio exceeds unity. It is the ratio of the larger number to the smaller. What does it mean? It means that in 68 cases out of 100 the actual value, if it exceed the estimate, will not exceed it by more than 12 per cent, and, if it fall below the estimate, will stay within a limit such that the estimate will not be more than 12 per cent greater than the actual value. This is not a convenient form, since this ratio always expresses the larger value in terms of the smaller value. It would be more convenient to have it always in terms of a percentage of the estimate. This may be done by putting $S_{\log y}$ in negative terms, and getting the corresponding natural value. The value $-.04901 = 9.95099 - 10$, which is the logarithm of .8933. In this form the ratio is based upon the relation of the smaller to the larger number. To make S_r readily intelligible we may combine the two, writing

$$S_r = .89 \text{ to } 1.12$$

Interpreting this, it means that, given a normal distribution, in 68 cases out of 100 the actual value will not be less than 89 per cent of the estimate, or more than 112 per cent of the estimate. This has a simple, definite meaning more

significant for most practical purposes than a similar measure in terms of absolute values.[1]

To find the values of $2S$ or $3S$ these percentage figures may not be simply multiplied by 2 or 3. The value of $S_{\log y}$ must be so multiplied, and the resulting values reduced to natural numbers. For convenience in use, the anti-logarithms of both the positive and negative values should be secured, as in the preceding case. The computations are simple.

$$2S_{\log y} = .09802$$

The anti-logarithm of this value, when considered positive is 1.25, when negative .80.

$$3S_{\log y} = .14703$$

The corresponding anti-logarithms are 1.40 and .71. Summarizing for the standard error, we have

$$S_r = .89 \text{ to } 1.12$$
$$2S_r = .80 \text{ to } 1.25$$
$$3S_r = .71 \text{ to } 1.40$$

The values given for S_r indicate the probable percentage limits within which actual value and estimated value should fall in 68 out of 100 cases. The values given for $2S_r$ indicate the probable percentage limits in 95 out of 100 cases. The values of $3S_r$ indicate the probable percentage limits in 99.7 cases out of 100, always on the assumption of a normal distribution of the logarithms of the actual values about the fitted curve.

Application of the Standard Error of Estimate

We may illustrate the use of $S_{\log y}$. Given a production of oats 50 per cent above the trend value (i.e. the ratio to

[1] The significance of a measure of reliability in percentage form was pointed out by D. H. Davenport in 1922, in an unpublished article, and such a measure has been employed in several studies. There has not been available, however, a ready method of computing this measure, and its possibilities have not, therefore, been fully realized.

trend is 1.50), what is the most probable accompanying price ratio and what is the degree of accuracy of this estimate?

The estimating equation is

$$\log Y = (9.99139 - 10) - 1.18206 \log X.$$

Substituting in this equation the value .176091 (the logarithm of 1.50) we secure for log Y the value 9.78324 − 10. The corresponding natural number is .607. This means that if production is 150 per cent of normal (as measured by the given line of trend) price will probably be 60.7 per cent of normal (as measured by the line of trend). To determine the reliability of this estimate, the standard error must be secured. Employing the values of S_r already computed we find that 54 is 89 per cent of 60.7, while 68 is 112 per cent of 60.7. We interpret these figures to mean that in 68 cases out of 100 the actual price prevailing under the given production conditions will not be less than 54 per cent of the normal or trend value nor more than 68 per cent of normal.[1] Corresponding values for $2S_r$ and $3S_r$ may be determined in the manner outlined above.

The Index of Correlation Based on Logarithmic Values

We have still to compute the third measure, the abstract index of correlation.[2] For an equation of the type

$$\log Y = \log a + b \log X$$

the formula for ρ reduces to

[1] A question arises at once as to the adequacy of the given lines of trend in thus determining the reliability of estimates in the present problem. This question is discussed in greater detail in another section.

[2] The symbol ρ is used for this measure of correlation, instead of r, even though the relationship in logarithmic form is linear. This is done because such a measure, in terms of logarithms, cannot be interpreted in precisely the same way as the ordinary coefficient of correlation.

$$\rho^2_{\log y \log x} = \frac{\log a\Sigma(\log Y) + b\Sigma(\log X \cdot \log Y) - Nc^2_{\log y}}{\Sigma(\log^2 Y) - Nc^2_{\log y}}$$

where $c_{\log y}$ represents the difference between the arithmetic mean of the logarithms of the Y-values and the origin (in this case, zero on the logarithmic scale). Substituting the proper values we have

$$\rho^2_{\log y \log x} = \frac{(-.00861 \times -.32849) + (-1.18206 \times -.1143005) - (33 \times .00009909)}{.21721807 - (33 \times .00009909)}$$

$$= \frac{.13466882}{.2139481}$$

$$= .629445$$

$$\rho_{\log y \log x} = .793$$

The index of correlation has a value of .793. How is this to be interpreted when we are dealing with logarithms as in the present case?

Its significance may be clearer if viewed in terms of the relationship

$$\rho^2_{\log y \log x} = 1 - \frac{S^2_{\log y}}{\sigma^2_{\log y}}$$

In the present case these values are

$$S_{\log y} = .04901$$

$$\sigma_{\log y} = .08052$$

When these values are squared and inserted in the above formula we have

$$\rho^2_{\log y \log x} = 1 - \frac{.002402417}{.00648328}$$

and

$$\rho_{\log y \log x} = .793$$

What does this value measure? We have seen that r and the more general index ρ are abstract measures of the degree of relationship between two variables, as this relationship is described by given functions. The value of ρ in a given case depends upon the variability about the

fitted line, in relation to the variability about the mean of the Y's. If the variability of estimates is materially reduced when the equation of regression is used as a basis for estimates, instead of the mean Y, the equation may be assumed to describe a significant relationship. The value of ρ depends thus upon the relation between the two quantities, S_y and σ_y.

In the cases dealt with in the preceding chapter the variability in each case was measured in terms of absolute deviations, and the value of ρ depended upon the relation between the two given measures of absolute variability. The sole difference in the present case is that we are working in terms of *logarithmic* or *ratio variability*, deviations being measured in terms of logarithms instead of natural numbers.

The index ρ must be interpreted in the light of this fact. Its value, as always, depends upon the relation between two measures of variability, S^2 and σ^2, but in the present instance these are expressed in terms of logarithms. In brief, the value of ρ depends upon the relation between the *ratio variability* about the fitted curve and the *ratio variability* about the geometric mean of the Y's. (It is the geometric mean of the Y's, because that is the value corresponding to the arithmetic mean of the Y logarithms.)

We have here a set of measures, therefore, which perform in the field of ratios precisely the same service as is performed in the field of natural numbers by S and ρ (in the linear case, r). These measures are secured in the same way as are S and ρ, except that the equation of relationship from which they are derived is one in which the dependent variable is $\log Y$ (or, in the reverse case, $\log X$). The general formulas for computing these values are the same as in dealing with natural numbers, except that $\log Y$ replaces Y throughout. The operation is analogous to that of using logarithmic paper instead of natural scale paper.

It should be noted that the values are in logarithmic or

ratio form if Y is expressed logarithmically, whether X be so expressed or not. Thus we have fitted a curve of the type

$$\log Y = \log a + b \log X$$

the logarithmic form of the ordinary parabola or hyperbola. The values S and ρ would also be in logarithmic form if the curve were of the type

$$\log Y = \log a + X \log b$$

the logarithmic form of the exponential

$$Y = a(b^X)$$

In each of these cases the logarithmic equation is linear, but this is not essential to the use of these measures. S and ρ are generally applicable measures, whether ratios or natural numbers be dealt with, and whether the functions be linear or otherwise.

It may be well at this point to summarize the symbols that have been used and to distinguish the different measures. We may employ the symbols S_y, σ_y and ρ when arithmetic relations are in question, the two former being measures of variation in absolute terms, and the index ρ referring to degree of relationship when natural numbers are employed. If the logarithms of the Y's are used it is advisable to distinguish the symbols by subscripts, using $S_{\log y}$ and $\sigma_{\log y}$ as measures of the logarithmic variation about the fitted curve and about the arithmetic mean of the logarithms of the Y's, respectively. If $S_{\log y}$ is reduced to ratio form, it may be written S_r. Since the index ρ must be interpreted somewhat differently in this case, it may be written $\rho_{\log y \log x}$, or $\rho_{\log y\, x}$.

The Use of Reciprocals in the Measurement of Relationship

Another type of curve may be used to describe the relationship between the production and price of oats, and

its use introduces us to a third field of correlation, a field in which somewhat new concepts enter, and in which the various measures must be interpreted in still another way. This is a curve of the type

$$Y = \frac{1}{a + bX}$$

which may be expanded by adding additional terms to the denominator, as

$$Y = \frac{1}{a + bX + cX^2}$$

This hyperbolic form has been used in several studies as an approximation to a "demand" curve for various commodities.

The equation to a curve of this type may be written

$$\frac{1}{Y} = a + bX$$

which is the equation to a straight line describing the relationship between the reciprocals of the Y's and the original X values. The normal equations required in fitting a curve of this type are

$$\text{I} \quad \Sigma\left(\frac{1}{Y}\right) = Na + b\Sigma(X)$$

$$\text{II} \quad \Sigma\left(\frac{X}{Y}\right) = a\Sigma(X) + b\Sigma(X)^2$$

The method of computing the necessary values is illustrated in Table 116.

Substituting the proper values in the normal equations, we have

$$34.3360320 = 33a + 33.338b$$

$$35.2571485 = 33.338a + 34.168554b$$

Solving,

$$a = -.1357$$

$$b = 1.1643$$

TABLE 116

Computation of Values Required in Fitting a Curve to Data of Oat Production and Prices

Example II

(1) *Year*	(2) *Price Ratio* Y	(3) *Production Ratio* X	(4) $\frac{1}{Y}$	(5) $\frac{X}{Y}$	(6) $\left(\frac{1}{Y}\right)^2$	(7) X^2
1881	1.30	.929	.7692308	.7146154	.59171602	.863041
1882	1.05	1.036	.9523810	.9866667	.90702957	1.073296
1883	.90	1.156	1.1111111	1.2844444	1.23456788	1.336336
1884	.85	1.128	1.1764706	1.3270588	1.38408307	1.272384
1885	.84	1.165	1.1904762	1.3869048	1.41723358	1.357225
1886	.77	1.108	1.2987013	1.4389610	1.68662507	1.227664
1887	.96	1.124	1.0416667	1.1708334	1.08506951	1.263376
1888	.79	1.151	1.2658228	1.4569620	1.60230736	1.324801
1889	.81	1.188	1.2345679	1.4666667	1.52415790	1.411344
1890	1.48	.798	.6756757	.5391892	.45653765	.636804
1891	1.10	1.088	.9090909	.9890909	.82644626	1.183744
1892	1.09	.943	.9174312	.8651376	.84168001	.889249
1893	1.16	.882	.8620690	.7603449	.74316296	.777924
1894	1.07	.886	.9345794	.8280373	.87343865	.784996
1895	.75	1.070	1.3333333	1.4266666	1.77777769	1.144900
1896	.76	.983	1.3157895	1.2934211	1.73130201	.966289
1897	.96	.969	1.0416667	1.0093750	1.08506951	.938961
1898	.95	1.005	1.0526316	1.0578948	1.10803329	1.010025
1899	.83	1.074	1.2048193	1.2939759	1.45158955	1.153476
1900	.86	1.033	1.1627907	1.2011628	1.35208221	1.067089
1901	1.37	.857	.7299270	.6255480	.53279343	.734449
1902	1.03	1.131	.9708738	1.0980583	.94259594	1.279161
1903	1.14	.911	.8771930	.7991228	.76946756	.829921
1904	.86	1.033	1.1627907	1.2011628	1.35208221	1.067089
1905	.86	1.090	1.1627907	1.2674419	1.35208221	1.188100
1906	1.04	1.013	.9615385	.9740385	.92455629	1.026169
1907	1.31	.770	.7633588	.5877863	.58271666	.592900
1908	1.29	.796	.7751938	.6170543	.60092543	.633616
1909	1.03	.978	.9708738	.9495146	.94259594	.956484
1910	.81	1.064	1.2345679	1.3135802	1.52415790	1.132096
1911	1.14	.810	.8771930	.7105263	.76946756	.656100
1912	.80	1.221	1.2500000	1.5262500	1.56250000	1.490841
1913	.87	.948	1.1494253	1.0896552	1.32117852	.898704
	32.83	33.338	34.3360320	35.2571485	36.85702940	34.168554

The desired equation is, therefore,

$$\frac{1}{Y} = -.1357 + 1.1643X$$

The Standard Error and the Index of Correlation in Terms of Reciprocals

To determine the utility of this equation we must have the standard error and the index of correlation. The two necessary formulas may be derived as in the preceding cases. Representing by $\frac{1}{Y}$ the reciprocal of an actual value we have, for each residual,

$$d = a + bX - \frac{1}{Y}$$

Multiplying by d and summing

$$\Sigma(d^2) = a\Sigma(d) + b\Sigma(dx) - \Sigma\left(\frac{d}{Y}\right)$$

Since

$$\Sigma(d) = 0 \text{ and } \Sigma(dX) = 0,$$

we have

$$\Sigma(d^2) = -\Sigma\left(\frac{d}{Y}\right)$$

Multiplying the residual equation now by $\frac{1}{Y}$, and summing, we have

$$\Sigma\left(\frac{d}{Y}\right) = a\Sigma\left(\frac{1}{Y}\right) + b\Sigma\left(\frac{X}{Y}\right) - \Sigma\left(\frac{1}{Y}\right)^2$$

Substituting the equivalent of $\Sigma\left(\frac{d}{Y}\right)$ in the preceding equation, we secure

$$\Sigma(d^2) = \Sigma\left(\frac{1}{Y}\right)^2 - a\Sigma\left(\frac{1}{Y}\right) - b\Sigma\left(\frac{X}{Y}\right)$$

and for $S_{\frac{1}{y}}^2$, we have

$$S_{\frac{1}{y}}^2 = \frac{\Sigma\left(\frac{1}{Y}\right)^2 - a\Sigma\left(\frac{1}{Y}\right) - b\Sigma\left(\frac{X}{Y}\right)}{N}$$

Inserting this value of $S_{\frac{1}{y}}^2$ in the general formula for the index of correlation

$$\rho^2 = 1 - \frac{S_{\frac{1}{y}}^2}{\sigma_{\frac{1}{y}}^2}$$

and simplifying, we have

$$\rho^2_{\frac{1}{y}x} = \frac{a\Sigma\left(\frac{1}{Y}\right) + b\Sigma\left(\frac{X}{Y}\right) - Nc_{\frac{1}{y}}^2}{\Sigma\left(\frac{1}{Y}\right)^2 - Nc_{\frac{1}{y}}^2}$$

Inserting the proper values in these two equations, we find that

$$S_{\frac{1}{y}} = .1191$$

$$\rho_{\frac{1}{y}x} = .766$$

For the standard deviation of the original Y-values, in terms of reciprocals, we secure

$$\sigma_{\frac{1}{y}} = .1851$$

(The subscript $\frac{1}{y}$ is used in connection with each of these measures, as they should be distinguished from measures based upon natural numbers or logarithms.)

Interpretation of the Standard Error of Estimate

How may we interpret these results? As in all former problems of this type the equation gives us a means of estimating Y from a known value of X. The standard error $S_{\frac{1}{y}}$ serves as a measure of the reliability of such estimates, and $\rho_{\frac{1}{y}x}$ is an abstract measure of the degree of relationship between the two variables. But in the present case all these measures are in terms of reciprocals. The equation enables us to estimate the reciprocal of Y, the

standard error has significance only in the form of a reciprocal, and the value of ρ depends upon the relation between two measures ($S_{\frac{1}{Y}}^2$ and $\sigma_{\frac{1}{Y}}^2$) both of which are in terms of reciprocals.

An illustration may make these meanings clear. If, in a given year, the production of oats is 150 per cent of normal, what is the most probable price? Substituting in the equation

$$\frac{1}{Y} = -.1357 + 1.1643X$$

a value of 1.50 for X, we have

$$\frac{1}{Y} = 1.6108$$

and

$$Y = .621$$

We may expect a price approximately 62 per cent of normal. As a measure of the reliability of this estimate, we have

$$S_{\frac{1}{Y}} = .1191$$

This must be applied to the estimate in terms of reciprocals. Thus we have

$$1.6108 + .1191 = 1.7299$$
$$1.6108 - .1191 = 1.4917$$

Reducing these reciprocals to natural numbers we secure .578 and .670 as the desired values. The most probable price, then, is 62.1 per cent of normal, and, on the assumption of an approximately normal distribution of reciprocals about the curve, the odds are 68 out of 100 that the price will fall between 57.8 per cent of normal and 67.0 per cent of normal. The limits of $2S$ and $3S$ may be similarly determined by adding to and subtracting from the estimate, as a reciprocal, amounts equal to twice .1191 and three times .1191. The results secured may then be converted to natural numbers. Just as with logarithms, the value

in absolute terms of a given difference between reciprocals varies at different points within the range of Y-values. Accordingly, the limits of reliability determined from $S_{\frac{1}{y}}$ should be expressed in natural numbers only after a particular estimate has been made.

A Comparison of Measures of Relationship Derived from Arithmetic, Geometric and Harmonic Measures

In interpreting ρ similar considerations enter. The value of the index of correlation, as we have seen, depends upon the degree of variation about the curve, as compared with the variation about the average of the original dependent series. In handling natural numbers, variability about the fitted line is compared with the variability about the *arithmetic mean* of the dependent variable, both measured in absolute terms (i.e. S_y is compared with σ_y). In handling logarithms, variability about the fitted line is compared with variability about the arithmetic mean of the logarithms of the dependent series, variability being measured in each case in terms of logarithms. But logarithmic deviations, as we have seen, may be interpreted in terms of ratios. The logarithmic deviations from the line represent the ratios of actual values to computed, while logarithmic deviations about the arithmetic mean of the logarithms of the original series represent the ratios of the actual values of the dependent series to their *geometric mean*. The value of $\rho_{\log y}$ depends upon the relation between these respective deviations (i.e. $S_{\log y}$ is compared with $\sigma_{\log y}$).

In fitting a curve in which the reciprocals of the dependent variable are employed, variability about the fitted line is measured in terms of reciprocals, and the variability of the original series is measured in the same terms. That is, $\sigma_{\frac{1}{y}}$ is computed from the differences between the recip-

rocals of the actual values and the arithmetic mean of all these reciprocals. But the arithmetic mean of these reciprocals is the reciprocal of the harmonic mean. Thus, in short, the value of the index of correlation, $\rho_{\frac{1}{y}}$, depends upon the relation between variability about the fitted line and variability about the *harmonic mean* of the dependent series, variation in both cases being measured in terms of reciprocals (i.e. $S_{\frac{1}{y}}$ is compared with $\sigma_{\frac{1}{y}}$).

We have, therefore, three broad families of curves for describing the relationship between variable quantities. These are:

1. Curves in the fitting of which natural values of the dependent variable are employed. Equations to all curves of this family will be of the type
$$Y = f(X)$$
2. Curves in the fitting of which logarithms of the dependent variable are employed. In all such cases the equations will be of the type
$$\log Y = f(X)$$
3. Curves in the fitting of which reciprocals of the dependent variable are employed. For these curves the equations will be of the type
$$\frac{1}{Y} = f(X)$$

In any one of these three cases the equations may be linear or non-linear. In so far as this problem of interpretation is concerned, there is no limitation as to the function of X which may be employed. (The computation of S and ρ by the methods suggested above involves certain limitations which are outlined elsewhere.)

The standard error of estimate for the first family of curves is derived in terms of the original units of measurement (for the dependent variable) and has a direct and simple meaning in these terms. The index of correlation,

for curves of this type, is a measure of the degree to which the *absolute variability* of the dependent variable may be lessened by measuring deviations from the fitted curve instead of from the *arithmetic mean.*

The standard error of estimate for the second family of curves is derived, by the method outlined, in terms of logarithms. It is more convenient in general to give it meaning in terms of *ratios.* The index of correlation, $\rho_{\log y}$, is a measure of the degree to which the logarithmic or ratio variability of the dependent variable may be lessened by computing deviations (or ratios) with the fitted curve instead of the *geometric mean* as base.

The standard error of estimate for the third family of curves is derived by the same process as in the other cases, but emerges as a *reciprocal.* The index of correlation, $\rho_{\frac{1}{y}}$, is a measure of the degree to which the variability of the dependent variable, *in terms of reciprocals*, may be lessened by computing reciprocal deviations from the fitted curve instead of from the *harmonic mean.*

Factors Governing the Choice of Measures of Relationship

It is clear, therefore, that the choice of a type of curve to describe a given relationship must be governed by basic considerations as to the type of average which is most appropriate as a measure of the central tendency of the given series. And this brings in a related question as to whether the dispersion about this average more nearly approximates the normal type when measured in absolute terms, in logarithms, or in reciprocals. In selecting a curve and in using the measures S and ρ there is always present an implicit assumption with respect to these points.

When absolute values are important, and the dispersion of the dependent variable approaches the normal type when plotted on an arithmetic scale, measures of relationship of

the arithmetic type would appear to be appropriate. But, as we have seen, in handling series in which *rates* of change rather than absolute amounts of change are of primary importance and the dispersion appears to follow a geometric law, the arithmetic mean and other arithmetic measures are notoriously inadequate. In such cases logarithmic curves seem preferable to arithmetic, and measures of the reliability of estimates and of degree of relationship which are based upon *ratios* seem to be more suitable than those based upon absolute values.

The harmonic mean has not been so widely employed as either of the above averages, and some attention may be given to principles governing its use in problems of the type here considered. In general, such harmonic measures are marked by the same weaknesses as the arithmetic, except that they err in the opposite direction. Geometric measures are perhaps better adapted to all-around employment than either. Yet in one particular field of interest to the economist the harmonic mean is particularly appropriate, and the utilization of reciprocals, as in the preceding example, seems to be justified.

The use of the harmonic mean assumes a normal distribution of reciprocals which, in natural numbers, means a much wider scatter above the average than below. The use of a curve of the type

$$\frac{1}{Y} = a + bX$$

involves a similar assumption as to the relation between Y and X. A given absolute increase in X will be accompanied by a certain decrease in the value of Y. The same absolute decrease in X will be accompanied by an increase in the value of Y which is larger than the decrease registered in the preceding case. But this is the relation which prevails, for many commodities, between the amounts produced and the price, the latter considered dependent. A given increase in production will cause some lowering of

price. An equal decrease will cause a much greater increase in price. Moreover, when averaging the prices of such commodities over a period, the harmonic mean may give a more typical value than any other average.[1] In such cases there is a strong *a priori* justification for using a curve of the reciprocal type and measuring the accuracy of all estimates in terms of harmonic relations.

A Comparison of Arithmetic, Geometric and Harmonic Measures

The contrast between these different methods may be brought home most effectively by comparing the results obtained when curves of these three types are fitted to the

[1] "Buyers and sellers of potatoes are frequently mistaken as to the price justified by fundamental economic conditions. If such an error is general in the fall, it may happen, for example, that the price which results is too high. If the price is too high in the early part of the season, potatoes will not be consumed fast enough to dispose of the supply available. Farmers and dealers will then find that not all of the stocks on hand can be sold at existing prices. Since potatoes can not be carried over from one year to the next, the price, under such conditions as have been mentioned, must be lowered enough to permit the supply to be disposed of before the end of the season. A properly adjusted price would remain the same throughout the season, except for a gradual advance to cover cost of storage, and would maintain a fairly uniform consumption throughout the season. But since an abnormally high price early in the season causes small consumption, it must be compensated by an abnormally low price during the remainder of the season, or not all the crop can be sold.

Similarly, if the price is abnormally low early in the season, the supply will be exhausted too rapidly and those who still have potatoes will find that they can get abnormally high prices for them during the remainder of the season."

But how, given the abnormally high or abnormally low prices during part of a season, may we compute the average price which would be justified by the true conditions of demand and supply, if these had been correctly estimated? Since "a low price during part of a season will be compensated only by a disproportionately high price during the remainder of the season" the arithmetic average for an entire season "will be somewhat higher than the average which would have resulted had a proper price been established at the beginning of the season. *This difficulty is eliminated by taking the harmonic mean of the monthly prices.*"

Holbrook Working, *Factors Determining the Price of Potatoes in St. Paul and Minneapolis.* Technical Bulletin 10, University of Minnesota Agricultural Experiment Station, 8–10.

same data. The computations involved in fitting curves of the second and third types (logarithmic and reciprocal) have been illustrated with reference to the data of oat production and prices (Table 114). A straight line (arithmetic) is fitted to the same data, and the necessary accompanying measures computed. The three sets of results are brought together in the following table:

TABLE 117

Relation Between the Production and Price of Oats
1881–1913

Comparison of Results of Curve Fitting
(Prices are the dependent variable in each case)

	Equation	*Standard Error of Estimate*	*Index of Correlation*
A	$Y = 2.24 - 1.236X$	$S_y = .12$	$r = .783$
B	$\frac{1}{Y} = -.1357 + 1.1643X$	$S_{\frac{1}{y}} = .1191$	$\rho_{\frac{1}{y}} = .766$
C	$\text{Log } Y = -.00861 - 1.18206 \log X$	$S_{\log y} = .04901$	$\rho_{\log y} = .793$

It is impossible to compare the three standard errors as they stand, since only the first one is in the original units of measurement (ratio of actual yield to normal). In the following table are given estimates, based on each of these equations, as to the most probable price (in terms of ratio to normal) which would accompany each of five different conditions of production. Each estimate is accompanied by a series of values which indicate the limits set by the standard error. Throughout, the values of the estimates plus and minus S, $2S$, and $3S$ are given, in order to indicate the probable scatter of actual values about the estimates. The different amounts of variation which may be expected about each of the three lines of relationship are measured by the actual differences between the estimates and the limiting cases. These differences are given in the columns headed Δ. All values in this table are comparable, being reduced to the original units (ratio of price to normal).

TABLE 118

Comparison of Price Estimates and of Standard Errors of Estimate Based on Three Equations Relating to the Production and Price of Oats

(1) Value of X (ratio of production to normal)	(2) Estimated value of Y (ratio of price to normal) from arithmetic equation (A)	(3) Limits of arithmetic estimate	(4) Δ	(5) Estimated value of Y from reciprocal equation (B)	(6) Limits of estimate, reciprocal	(7) Δ	(8) Estimated value of Y from logarithmic equation (C)	(9) Limits of logarithmic estimate	(10) Δ
		+3S = 1.982	+.36		+3S = 11.223	+8.983		+3S = 3.114	+.890
		+2S = 1.862	+.24		+2S = 4.803	+2.563		+2S = 2.780	+.556
.5	1.622	+S = 1.742	+.12	2.240	+S = 3.055	+.815	2.224	+S = 2.491	+.267
		−S = 1.502	−.12		−S = 1.768	−.472		−S = 1.979	−.245
		−2S = 1.382	−.24		−2S = 1.461	−.779		−2S = 1.779	−.445
		−3S = 1.262	−.36		−3S = 1.244	−.996		−3S = 1.579	−.645
		+3S = 1.611	+.36		+3S = 2.281	+1.024		+3S = 1.786	+.510
		+2S = 1.491	+.24		+2S = 1.794	+.537		+2S = 1.595	+.319
.8	1.251	+S = 1.371	+.12	1.257	+S = 1.478	+.221	1.276	+S = 1.429	+.153
		−S = 1.131	−.12		−S = 1.093	−.164		−S = 1.136	−.140
		−2S = 1.011	−.24		−2S = .967	−.290		−2S = 1.021	−.255
		−3S = .891	−.36		−3S = .867	−.390		−3S = .906	−.370
		+3S = 1.364	+.36		+3S = 1.490	+.518		+3S = 1.372	+.392
		+2S = 1.244	+.24		+2S = 1.265	+.293		+2S = 1.225	+.245
1.0	1.004	+S = 1.124	+.12	.972	+S = 1.100	+.128	.980	+S = 1.098	+.118
		−S = .884	−.12		−S = .871	−.101		−S = .872	−.108
		−2S = .764	−.24		−2S = .789	−.183		−2S = .784	−.196
		−3S = .644	−.36		−3S = .722	−.250		−3S = .696	−.284
		+3S = 1.117	+.36		+3S = 1.106	+.313		+3S = 1.106	+.316
		+2S = .997	+.24		+2S = .977	+.184		+2S = .987	+.197
1.2	.757	+S = .877	+.12	.793	+S = .875	+.082	.790	+S = .885	+.095
		−S = .637	−.12		−S = .724	−.069		−S = .703	−.087
		−2S = .517	−.24		−2S = .667	−.126		−2S = .632	−.158
		−3S = .397	−.36		−3S = .618	−.175		−3S = .561	−.229
		+3S = .746	+.36		+3S = .798	+.177		+3S = .852	+.245
		+2S = .626	+.24		+2S = .728	+.107		+2S = .761	+.154
1.5	.386	+S = .506	+.12	.621	+S = .670	+.049	.607	+S = .680	+.073
		−S = .266	−.12		−S = .578	−.043		−S = .542	−.065
		−2S = .146	−.24		−2S = .541	−.080		−2S = .484	−.123
		−3S = .026	−.36		−3S = .508	−.113		−3S = .433	−.174

ZONES OF ESTIMATE AND THEIR SIGNIFICANCE

A careful study of this table should make clear the nature of estimates based on the three types of equations here

presented. The fundamental differences lie not so much in the actual values of the estimates, as in the standard errors which measure the reliability of these estimates and indicate the limits within which the actual values are likely to fall. In other words, the differences lie in the assumptions made as to the character of the scatter about the curves.

The measure S_y, which relates to the arithmetic curve, gives the same absolute range to errors of estimate whether the estimated value be high or low. An arithmetic dispersion about the curve is assumed. In each case the estimate is the arithmetic mean of the value which exceeds the estimate by an amount equal to S_y (or any multiple of S_y) and the value which falls below it by an equal amount. These conditions are brought out graphically in Fig. 81. The original points are plotted, the straight line of relationship (arithmetic) is shown, and zones of estimate having widths, respectively, of $2S$, $4S$ and $6S$, centering at the fitted line, are marked out.

The measure $S_{\log y}$ gives the same relative or percentage range to errors of estimate, whether the estimate be high or low. This means that the absolute range within which the actual values should fall is much less when the estimates are low than when they are high. It assumes a geometric dispersion about the curve which describes the relationship. The estimate is, in this case, the geometric mean of the value which exceeds it by an amount equal to $S_{\log y}$ (or any multiple of $S_{\log y}$) and the value which falls below it by an equal amount. Fig. 82 presents these relationships graphically. The original data are here plotted, together with the graph of the equation

$$Y = .9804X^{-1.18206}$$

There are shown, also, the limits of zones of estimate having widths equal, respectively, to $2Sr$, $4Sr$ and $6Sr$, centering (geometrically) at the line of relationship. A comparison of Fig. 81 and Fig. 82 will perhaps give a better understand-

ing of the differences between estimates based on the assumption of an arithmetic distribution and those based on the assumption of a geometric distribution.

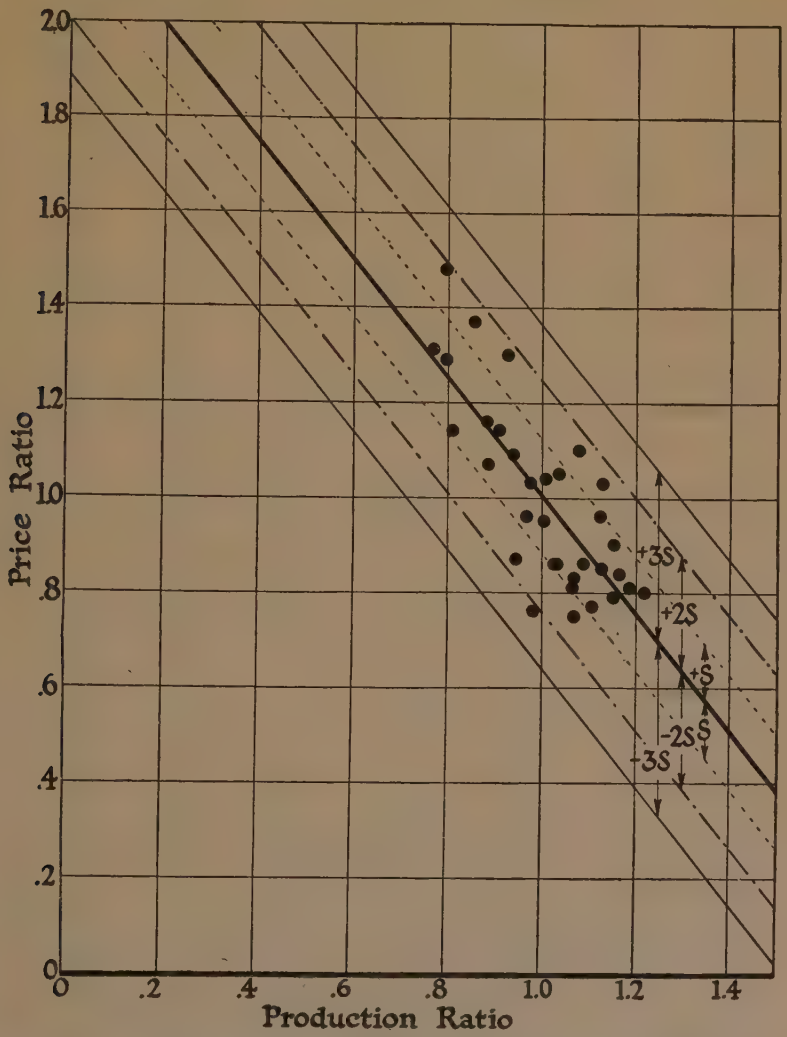


FIG. 81. — The Relation between the Production and Price of Oats; Illustrating the Use of an Arithmetic Equation of Regression and Arithmetic Zones of Estimate

The points and lines shown in Fig. 82 are plotted on a logarithmic scale in Fig. 83. On this scale the curve of relationship becomes straight, and the zones of estimate

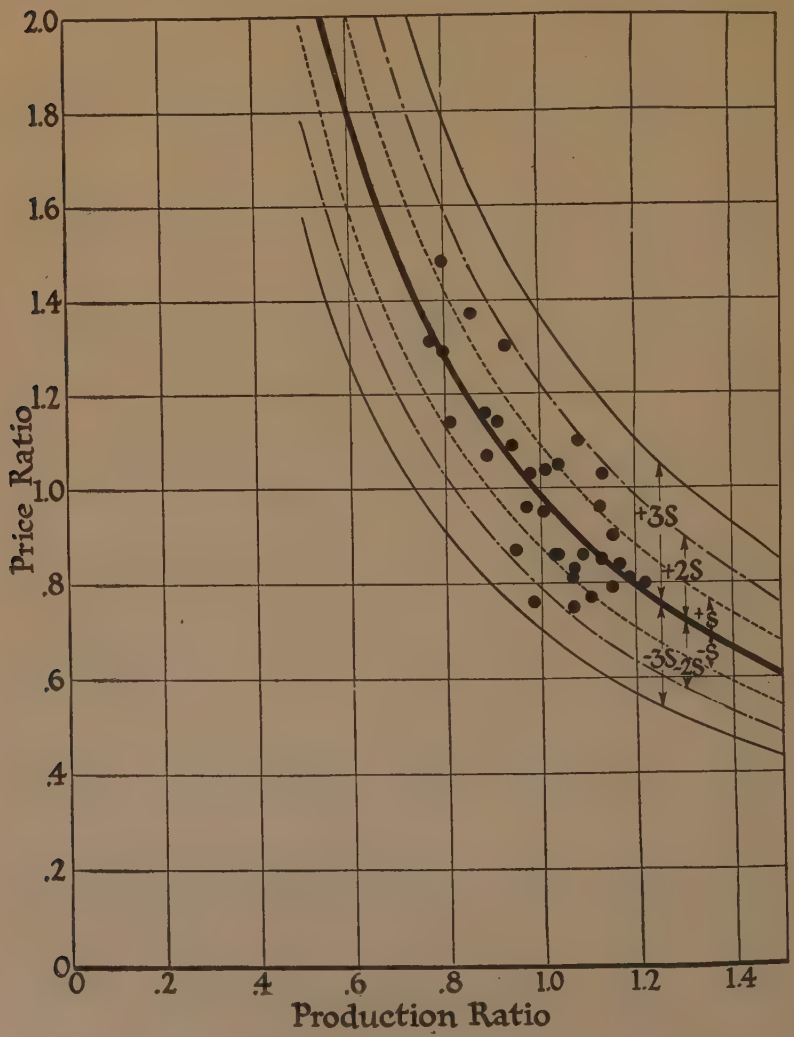


FIG. 82. — The Relation between the Production and Price of Oats; Illustrating the Use of a Logarithmic Equation of Regression and Geometric Zones of Estimate

appear as symmetrical and of equal width throughout the range. This transformation when the data are plotted on logarithmic paper makes clear the fundamental simplicity

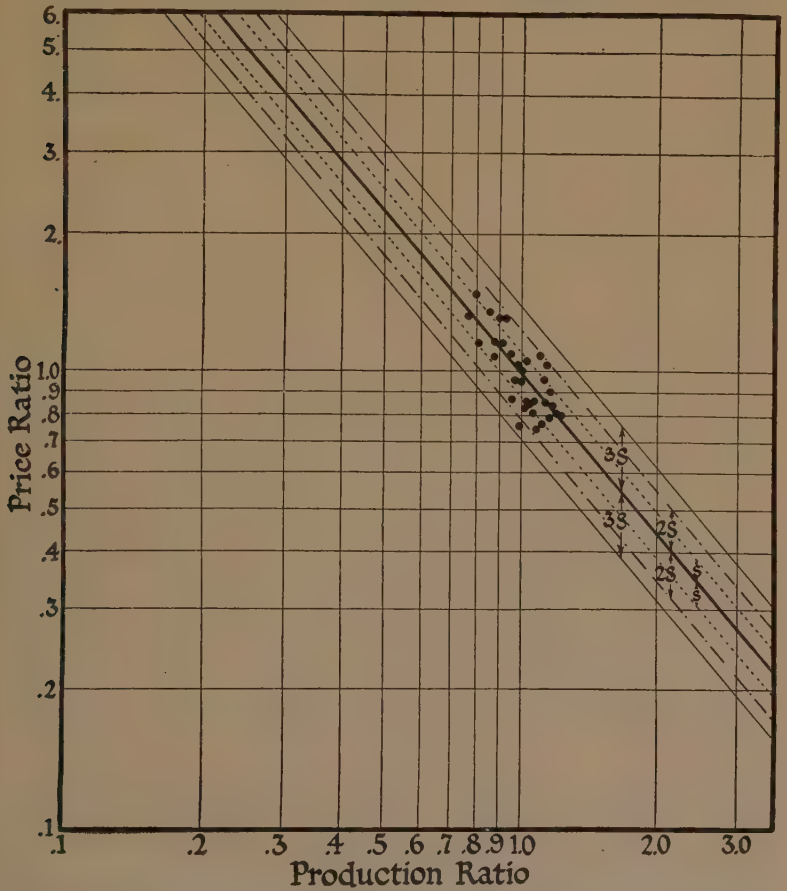


FIG. 83. — The Relation between the Production and Price of Oats; Illustrating the Use of a Logarithmic Equation of Regression and Geometric Zones of Estimate (Plotted on Double Logarithmic Paper)

of the assumptions involved in making estimates from logarithmic values.

In using the measure $S_{\frac{1}{y}}$ we carry still further the assumption that the variability about the curve is greater with

high prices than with low. It shows a very limited range to errors of estimate when the estimate is low and a very wide range when the estimated price is high. A harmonic dispersion about the curve is assumed. The computed value, or estimate, is always the harmonic mean of the value which exceeds it by an amount equal to $S_{\frac{1}{y}}$ (or any multiple of $S_{\frac{1}{y}}$) and the value which falls below it by an equal amount.

In Fig. 84 the curve $\frac{1}{Y} = -.1357 + 1.1643\,X$ is plotted, together with the original observations. Zones of estimate with widths of $2S_{\frac{1}{y}}$, $4S_{\frac{1}{y}}$ and $6S_{\frac{1}{y}}$, centering (harmonically) at the fitted line, are shown. The differences between this figure and each of the two preceding are quite marked, particularly with respect to the zones of estimate. On the assumption of a normal harmonic distribution about the curve describing the relationship, the outer zone (with width equal to $6S$) marks the limits within which 99.7 per cent of all the points should fall, and the inner zone (with width equal to $2S$) marks the limits within which 68 per cent of all the points should fall. By plotting reciprocals throughout, instead of natural numbers, this apparently abnormal distribution could be reduced to the symmetrical form secured in plotting the geometric values on the logarithmic chart.

For both high and low estimates the geometric measure, $S_{\log y}$, stands between the arithmetic measure, S_y, and the harmonic measure, $S_{\frac{1}{y}}$. While the two latter have their particular functions, and are appropriate in certain cases, it is probably true that in using such methods as these in economic analysis, measures of the geometric family are more generally useful than those of the other types. This means, merely, that ratios are usually more important than absolute differences. It seems reasonable therefore to base estimates upon an equation of the type

$$\text{Log } Y = f(X)$$

and to measure the reliability of these estimates in terms of ratios, using $S_{\log y}$. In such cases, as we have seen,

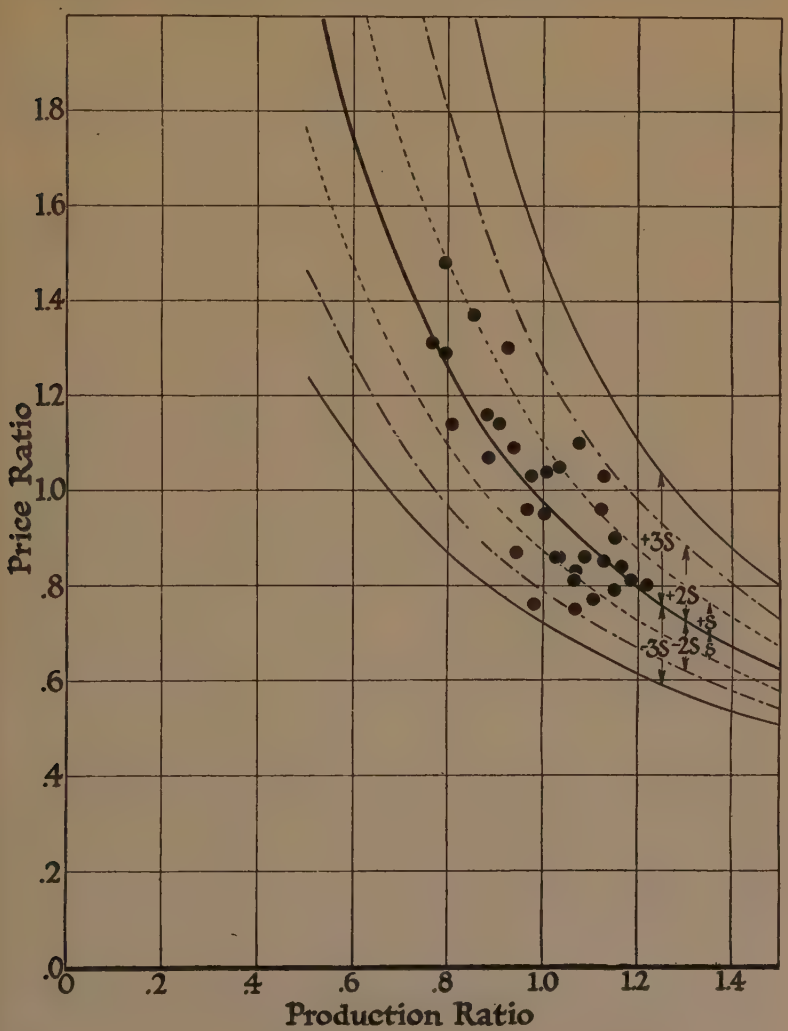


FIG. 84. — The Relation between the Production and Price of Oats; Illustrating the Use of an Equation of Regression Based upon Reciprocals and of Harmonic Zones of Estimate

correlation is measured by $\rho_{\log y}$, the value of which depends upon the *ratio variability* about the curve as compared with the *ratio variability* about the geometric mean.[1]

REFERENCES

KILLOUGH, H. B. *A Statistical Analysis of Oat Prices.* A manuscript prepared for the U. S. Bureau of Agricultural Economics.

MOORE, H. L. *Elasticity of Demand and Flexibility of Prices.* Journal of the American Statistical Association, March, 1922. *Empirical Laws of Demand and Supply and the Flexibility of Prices.* Political Science Quarterly, Dec., 1919.

WALSH, C. M. *The Problem of Estimation* (1–67).

Examples of curves employed to describe relationships of the type discussed in this chapter will be found in

SCHULTZ, HENRY. *The Statistical Measurement of the Elasticity of Demand for Beef.* Journal of Farm Economics, July, 1924 (254–278).

WORKING, HOLBROOK. *Factors Determining the Price of Potatoes in St. Paul and Minneapolis.* Technical Bulletin 10, University of Minnesota Agricultural Experiment Station.

[1] The reasoning in C. M. Walsh's book, *The Problem of Estimation*, (London, King, 1921) is peculiarly applicable to the present problem. Citing Galileo, in defence of the use of the geometric mean in averaging estimates, Walsh writes: "And so errors must be measured by an error *which is a ratio between the estimate and the true quantity, and not a concrete quantity itself.* We cannot measure errors by so many pounds, feet or crowns; we must measure them by the proportions of the pounds, feet or crowns in the erroneous estimates to the pounds, feet or crowns in the thing estimated." (Italics mine.) (p. 12.) This argument bears out powerfully what has been said as to the use of logarithmic functions in estimating, and as to the employment of logarithmic measures of errors of estimate.

CHAPTER XIV

THE MEASUREMENT OF RELATIONSHIP: MULTIPLE AND PARTIAL CORRELATION

In dealing with the problem of correlation in the preceding chapters we have been concerned with problems involving only two variables, a dependent variable and a single independent variable. We have found, in certain cases, a fairly high degree of correlation between the two variables studied. But it is obvious that, in general, economic phenomena are affected by more than one factor, that the fluctuations in a single variable may be due to the interaction of many forces. In dealing with just two variables all other factors are ignored, on the assumption, usually, that in the single independent variable are found the most important causes[1] of fluctuations in the dependent variable. Thus, in the alfalfa example given, the effect upon yield of but a single factor, irrigation, was studied. Yet variations in rainfall and temperature must have affected the yield in the different years studied. Similarly, variations in practically every factor dealt with in economic analysis are traceable to more than one cause. If our analysis is to be complete we must employ methods which will enable more than two variables to be handled at a time. We need instruments which will assist us in measuring the combined effect upon a single variable of a number of factors. Such instruments may be secured by a simple extension of methods already familiar.

In the following table are presented figures showing the yield of corn, per acre, in Kansas from 1890 to 1922,

[1] This should not be taken to mean that the coefficient of correlation measures or establishes causal relationships. Cf. Chapter XVI on this point.

TABLE 119

Corn Yield and Temperature in Kansas[1] 1890–1922

(1) Year	(2) Actual Yield per acre, in bushels	(3) Trend Value, Yield per acre, in bushels[2]	(4) Actual Yield as percentage of Trend Value X_1	(5) Average June Temperature X_2	(6) Average July Temperature X_3	(7) Average August Temperature X_4
1890	15.6	22.4	69.6	77.6	83.1	76.1
1891	26.7	22.2	120.3	70.7	74.0	75.1
1892	24.5	22.1	110.9	73.4	77.5	76.5
1893	21.3	21.9	97.3	74.7	79.5	73.8
1894	11.2	21.8	51.4	74.2	77.8	78.0
1895	24.3	21.6	112.5	71.7	74.9	76.0
1896	28.0	21.5	130.2	74.1	78.1	78.7
1897	18.0	21.3	84.5	76.6	80.2	76.0
1898	16.0	21.2	75.5	75.0	77.7	78.2
1899	27.0	21.0	128.6	73.9	76.2	80.6
1900	19.0	20.9	90.9	74.9	77.9	81.0
1901	7.8	20.7	37.7	77.3	85.0	79.1
1902	29.9	20.6	145.1	70.9	76.8	78.2
1903	25.6	20.4	125.5	67.2	78.3	75.3
1904	20.9	20.3	103.0	70.4	75.6	74.6
1905	27.7	20.1	137.8	75.5	74.5	78.7
1906	28.9	20.0	144.5	71.8	73.8	76.3
1907	22.1	19.8	111.6	72.0	78.4	78.1
1908	22.0	19.7	111.7	72.1	75.8	76.2
1909	19.9	19.5	102.1	73.1	78.1	80.1
1910	19.0	19.4	97.9	72.2	79.5	75.7
1911	14.5	19.2	75.5	80.5	78.6	76.4
1912	23.0	19.1	120.4	69.3	79.9	77.4
1913	3.2	18.9	16.9	74.2	82.1	84.2
1914	18.5	18.8	98.4	78.2	79.9	78.2
1915	31.0	18.6	166.7	69.2	74.0	70.1
1916	10.0	18.5	54.1	70.3	81.2	79.6
1917	13.0	18.3	71.0	72.8	80.8	73.4
1918	7.1	18.2	39.0	78.4	78.3	82.3
1919	15.2	18.0	84.4	72.3	80.2	78.3
1920	26.5	17.9	148.0	72.8	77.6	72.9
1921	22.2	17.7	125.4	74.4	79.2	78.6
1922	19.3	17.6	109.7	75.2	77.0	80.1

[1] The data of corn yield are from *Bulletin* 515, U. S. D. A., and from the *Yearbooks* of the U. S. D. A, Temperature data are from reports of the U. S. Weather Bureau.

[2] The equation to the line of trend is $Y = 19.967 - .1502X$, with origin at 1906.

together with the average June, July, and August temperatures for each of these years. A straight line of trend has been fitted to the yield figures, and the actual yield is given, in column (4), in percentages of normal, as measured by the line of trend.

The Relation between Corn Yield and Temperature; Preliminary Analysis

It is known that corn yield is affected by the temperature during the growing season. The object of the present study is the determination of the precise relation between yield and temperature during each of the three months given, in order to secure a basis for estimating the yield from a knowledge of the temperature. As certain growing months are more important than others, the relation between temperature and yield may be determined, first, for each of the three months separately.

The equation which describes the relationship between yield per acre and June temperature will be of the type

$$X_1 = a + b_{12}X_2$$

The equation describing the relationship between yield per acre and July temperature will be of the type

$$X_1 = a + b_{13}X_3$$

(In each case X_1 represents the percentage relation of actual to trend value, while X_2, X_3, etc., represent the absolute temperature, in degrees Fahrenheit.) Instead of using to represent the variables the symbols Y and X, as in the preceding examples, X_1, X_2, X_3, etc., are employed, X_1 representing in this case the dependent variable. The symbol for the constant representing the slope (the coefficient of regression) is, in the first instance above, b_{12}. The subscripts 1 and 2 indicate the variables to which this constant refers, the first subscript always representing the dependent variable (X_1 in the first example above), the

second the independent variable (X_2 in the present case). These subscripts are necessary to distinguish the different constants when several variables enter into the problem. The meaning is precisely the same as in the former examples when no subscripts were needed because only two variables were dealt with.

Solving the proper normal equations for the constants in the equation which describes the average relationship between yield per acre (percentage relation of actual to trend) and June temperature, we have

$$X_1 = 522.31 - 5.743X_2$$

The value of S_{12} may be determined from the formula

$$S_{12}^2 = \frac{\Sigma(X_1^2) - a\Sigma(X_1) - b_{12}\Sigma(X_1X_2)}{N}$$

(The subscripts to S, and those to r which appear below, have the same meaning as those employed in the preceding paragraph.) Substituting the given values, and solving, we have

$$S_{12}^2 = 913.178$$

and

$$S_{12} = 30.22$$

The significance of the standard error, S, as a measure of the reliability of estimates based upon the equation of relationship, has been fully explained. In judging of the usefulness of the equation, S_{12} should be compared with σ_1 (the standard deviation of X_1) which may be looked upon as a measure of the reliability of estimates based upon the arithmetic mean of the variable X_1. For this we have

$$\sigma_1 = 34.477$$

Clearly, the estimates from the equation are more reliable than those based upon the mean. The coefficient of correlation, r, expresses this relationship in abstract terms. We may get this value from the equation

$$r^2_{12} = \frac{a\Sigma(X_1) + b_{12}\Sigma(X_1X_2) - Nc_1^2}{\Sigma(X_1^2) - Nc_1^2}$$

Solving for r, and giving it the sign of b_{12}, we have

$$r_{12} = -.4814$$

These values indicate a negative correlation, though not a high one, between yield per acre of corn and June temperature in Kansas. Let us see if the estimates could be improved if based upon the temperature in July instead of in June.

The values needed in this study may be computed from Table 119. Solving for the constants in the equation of regression, we secure the equation

$$X_1 = 827.64 - 9.302X_3$$

For the standard error, we have

$$S_{13} = 24.73$$

and for the coefficient of correlation

$$r_{13} = -.6968$$

We have here a closer relation and a better basis for estimates than in the case when June temperature was considered.

Repeating the process for yield per acre and August temperature, we have

$$X_1 = 517.86 - 6.098X_4$$
$$S_{14} = 29.98$$
$$r_{14} = -.4937$$

August temperature, it is evident, also affects the corn yield in Kansas, a low temperature conducing to yield above normal. The relationship is not so close as in the case of July temperature, but it is still significant. What is needed now is some method of combining these three factors, in order that an estimate may be based upon a knowledge of the influence of these factors, in combination, upon the yield of corn. The addition or averaging of the temperatures in the three months will not do, for July is obviously more important than either of the other months. The principle of the method by which this may be accomplished is simple.

The Estimation of Corn Yield from Three Independent Variables

The estimating or regression equation in the present case will be one in which there is a single dependent variable (corn yield) and three independent variables. It will be of the form

$$X_1 = a + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$$

If we can determine the values of the four constants, we may substitute given values of X_2, X_3 and X_4 in the equation and thus get an estimate for X_1 in precisely the same way as when two variables are dealt with. The method of least squares affords the means of solving for the required constants.

The symbols require a word of explanation, as a perfectly simple equation is given a rather ponderous appearance by all the subscripts employed. The symbol b_{12}, it has been explained, represents the coefficient of regression of X_1 on X_2 (i.e. the slope of the line describing their relationship, X_1 being dependent) when these two variables alone are included in the study. The symbol $b_{12.34}$ represents the *coefficient of net regression* of X_1 on X_2. The addition of the subscripts 3 and 4 to the right of the period means, simply, that the variables X_3 and X_4 have been included in the study and their effect eliminated, in so far as this one constant ($b_{12.34}$) is concerned. This constant measures the weight which must be given to the variable X_2 in an estimate of X_1 based upon the three independent variables, X_2, X_3 and X_4. It will not, of course, be the same as b_{12}, which indicates the weight given to X_2 when an estimate of X_1 is based upon X_2 alone. Similarly the constant $b_{12.43}$, the coefficient of net regression of X_1 on X_3, measures the weight given to X_3 when X_2 and X_4 are also included. Each coefficient represents a single, simple constant, but the subscripts are necessary in order that the precise meaning of this constant may be clear. The sub-

scripts to the left of the period are termed *primary subscripts*, those to the right *secondary subscripts*.

Formation and Solution of the Normal Equations

The first task[1] is the securing of the normal equations required in solving for the constants in the estimating equation given above. Following the usual procedure[2] we have:

$$\text{I} \quad \Sigma(X_1) = Na + b_{12.34}\Sigma(X_2) + b_{13.24}\Sigma(X_3) + b_{14.23}\Sigma(X_4)$$

$$\text{II} \quad \Sigma(X_1X_2) = a\Sigma(X_2) + b_{12.34}\Sigma(X_2^2) + b_{13.24}\Sigma(X_2X_3) + b_{14.23}\Sigma(X_2X_4)$$

$$\text{III} \quad \Sigma(X_1X_3) = a\Sigma(X_3) + b_{12.34}\Sigma(X_2X_3) + b_{13.24}\Sigma(X_3^2) + b_{14.23}\Sigma(X_3X_4)$$

$$\text{IV} \quad \Sigma(X_1X_4) = a\Sigma(X_4) + b_{12.34}\Sigma(X_2X_4) + b_{13.24}\Sigma(X_3X_4) + b_{14.23}\Sigma(X_4^2)$$

The given values might be substituted in these simultaneous equations and solutions secured directly for the four constants. It is possible to reduce the number of normal equations by one, however, and thus lessen materially the labor of computation. This is done by using deviations from the arithmetic mean for each variable instead of absolute values, getting rid in this way of the constant term a in the original equation.

If we let A_1, A_2, A_3, etc., represent the arithmetic means of the different variables while x_1, x_2, x_3, etc., represent deviations from the means, we may replace the absolute numbers X_1, X_2, X_3, etc., by their equivalents, $x_1 + A_1$,

[1] The approach to the problem of multiple correlation which is here taken follows that of H. R. Tolley and M. J. B. Ezekiel ("A Method of Handling Multiple Correlation Problems," *Journal of the American Statistical Association*, December, 1923, 993–1003). Statisticians are indebted to Messrs. Tolley and Ezekiel for the simple and effective method they have evolved. For details not here given, relating to the simplification of the normal equations, readers are referred to this article.

[2] *Cf.* Appendix A for a discussion of this procedure.

$x_2 + A_2$, $x_3 + A_3$, etc. Making these substitutions in the normal equations, certain algebraic simplifications are possible which eliminate the first of the normal equations, and reduce the others to the following form:

$$\frac{\Sigma(x_1x_2)}{N} = \frac{\Sigma(x_2^2)}{N} b_{12.34} + \frac{\Sigma(x_2x_3)}{N} b_{13.24} + \frac{\Sigma(x_2x_4)}{N} b_{14.23}$$

$$\frac{\Sigma(x_1x_3)}{N} = \frac{\Sigma(x_2x_3)}{N} b_{12.34} + \frac{\Sigma(x_3^2)}{N} b_{13.24} + \frac{\Sigma(x_3x_4)}{N} b_{14.23}$$

$$\frac{\Sigma(x_1x_4)}{N} = \frac{\Sigma(x_2x_4)}{N} b_{12.34} + \frac{\Sigma(x_3x_4)}{N} b_{13.24} + \frac{\Sigma(x_4^2)}{N} b_{14.23}$$

All the variables in the above equations refer to deviations from the respective arithmetic means. Therefore $\frac{\Sigma(x_1x_2)}{N}$ is simply the mean product of the variables x_1 and x_2, $\frac{\Sigma(x_2^2)}{N}$ is σ_2^2, etc. Representing the various mean products by the symbols p_{12}, p_{13}, etc., and inserting the symbols for the squares of the standard deviations, we secure, for the normal equations:

$$p_{12} = \sigma_2^2 b_{12.34} + p_{23} b_{13.24} + p_{24} b_{14.23}$$
$$p_{13} = p_{23} b_{12.34} + \sigma_3^2 b_{13.34} + p_{34} b_{14.23}$$
$$p_{14} = p_{24} b_{12.34} + p_{34} b_{13.24} + \sigma_4^2 b_{14.23}$$

This is the most convenient form for the solution of the normal equations.

From the data, as arranged in Table 119, the following values are derived:

$$\Sigma(X_1) = 3298.1 \qquad \Sigma(X_1^2) = 368846.67$$
$$\Sigma(X_2) = 2426.9 \qquad \Sigma(X_2^2) = 178755.75$$
$$\Sigma(X_3) = 2581.5 \qquad \Sigma(X_3^2) = 202163.79$$
$$\Sigma(X_4) = 2553.8 \qquad \Sigma(X_4^2) = 197890.32$$

$$\Sigma(X_1X_2) = 240967.22$$
$$\Sigma(X_1X_3) = 255954.11$$
$$\Sigma(X_1X_4) = 253664.85$$
$$\Sigma(X_2X_3) = 189941.83$$

$$\Sigma(X_2X_4) = 187909.38$$
$$\Sigma(X_3X_4) = 199845.00$$

$$c_1 = \frac{\Sigma(X_1)}{N}$$

$$= 99.9424 \qquad c_1^2 = 9988.4833$$
$$c_2 = 73.5424 \qquad c_2^2 = 5408.4846$$
$$c_3 = 78.2273 \qquad c_3^2 = 6119.5105$$
$$c_4 = 77.3879 \qquad c_4^2 = 5988.8871$$

From these values, the quantities necessary for the solution of the normal equations may be readily determined. These quantities are brought together below:

$$\sigma_1^2 = \frac{\Sigma(X_1^2)}{N} - c_1^2$$

$$= \frac{368,846.67}{33} - 9988.4833 = 1188.688$$

$$\sigma_2^2 = \frac{178,755.75}{33} - 5408.4846 = 8.3564$$

$$\sigma_3^2 = \frac{202,163.79}{33} - 6119.5105 = 6.6645$$

$$\sigma_4^2 = \frac{197,890.32}{33} - 5988.8871 = 7.7893$$

$$p_{12} = \frac{\Sigma(X_1X_2)}{N} - c_1c_2$$

$$p_{12} = \frac{240,967.22}{33} - (99.9424 \times 73.5424) = -47.967$$

$$p_{13} = \frac{255,954.11}{33} - (99.9424 \times 78.2273) = -62.039$$

$$p_{14} = \frac{253,664.85}{33} - (99.9424 \times 77.3879) = -47.519$$

$$p_{23} = \frac{189,941.83}{33} - (73.5424 \times 78.2273) = 2.790$$

$$p_{24} = \frac{187,909.38}{33} - (73.5424 \times 77.3879) = 2.932$$

$$p_{34} = \frac{199,845.00}{33} - (78.2273 \times 77.3879) = 2.063$$

Substituting in the normal equations, we have:

$$-47.967 = 8.3564b_{12.34} + 2.790b_{13.24} + 2.932b_{14.23}$$
$$-62.039 = 2.790b_{12.34} + 6.6645b_{13.24} + 2.063b_{14.23}$$
$$-47.519 = 2.932b_{12.34} + 2.063b_{13.24} + 7.7893b_{14.23}$$

Solving these simultaneous equations[1] we secure the following values for the constants:

$$b_{12.34} = -2.095 \qquad b_{13.24} = -7.394 \qquad b_{14.23} = -3.354$$

The required equation is, therefore,

$$x_1 = -2.095x_2 - 7.394x_3 - 3.354x_4$$

This is the equation of net regression of x_1 on x_2, x_3 and x_4. Any given values of the three independent variables (June temperature, July temperature and August temperature) may be substituted in this equation, and the most probable value of the dependent variable (corn yield per acre) determined. In the equation as it stands, it should be noted, all the variables are expressed as deviations from their respective arithmetic means. For practical purposes it is advisable to have an equation in terms of the original values. In other words, it is desirable to shift the origin from the point of averages to the zero point on the original scales. This necessitates re-introducing the constant term a.

The value of a may be determined from the equation

$$A_1 = a + A_2b_{12.34} + A_3b_{13.24} + A_4b_{14.23}$$

where the large A's represent the respective arithmetic means.[2] Inserting the proper values, we have

[1] Any method of solution may be employed. Perhaps the most convenient with three or more equations is the Doolittle method. This is explained in detail in Appendix A.

[2] This equation is derived from the first normal equation, as given on p. 491,

$$\Sigma(X_1) = Na + b_{12.34}\Sigma(X_2) + b_{13.24}\Sigma(X_3) + b_{14.23}\Sigma(X_4)$$

Replacing the absolute numbers X_1, X_2, etc., by their equivalents, $x_1 + A_1$, $x_2 + A_2$, etc., we secure

$$\Sigma(x_1) + NA_1 = Na + b_{12.34}[\Sigma(x_2) + NA_2] + b_{13.24}[\Sigma(x_3) + NA_3] + b_{14.23}[\Sigma(x_4) + NA_4]$$

Since $\Sigma(x_1) = 0$, $\Sigma(x_2) = 0$, etc., these variables disappear. Dividing through by N we obtain the equation presented above.

$$99.9424 = a + (73.5424 \times -2.094816) + (78.2273 \times -7.39374) + (77.3879 \times -3.353796)$$ [1]

Solving,

$$a = 1091.94$$

The equation of regression in terms of original values is, therefore,

$$X_1 = 1091.94 - 2.095X_2 - 7.394X_3 - 3.354X_4$$

Computation of the Standard Error of Estimate

Are estimates based upon this equation any more reliable than those based upon the equations previously derived, each of which referred to a single independent variable? To answer this question the value of the standard error must be computed. This will be represented in the present case by $S_{1.234}$, the subscripts referring to the single dependent variable (X_1) and the three independent variables. This value may be computed from the formula [2]

$$S^2_{1.234} = \sigma_1^2 - b_{12.34}p_{12} - b_{13.24}p_{13} - b_{14.23}p_{14}$$

[1] The arbitrary origin is at zero on each of the original scales, hence $A_1 = c_1$, $A_2 = c_2$, etc. To ensure greater accuracy in solving for a, the values of the coefficients $b_{12.34}$, $b_{13.24}$, etc., are given to a greater number of decimal places than in the equation of regression.

[2] This formula may be derived as follows: Given an equation of the type

$$x_1 = b_{12.34}x_2 + b_{13.24}x_3 + b_{14.23}x_4$$

(in which the variables refer to deviations from the means) each residual may be computed from the equation

$$d = b_{12.34}x_2 + b_{13.24}x_3 + b_{14.23}x_4 - x_1 \quad \ldots\ldots (1)$$

Multiplying throughout by d, and adding, we have

$$\Sigma(d^2) = b_{12.34}\Sigma(dx_2) + b_{13.24}\Sigma(dx_3) + b_{14.23}\Sigma(dx_4) - \Sigma(dx_1)$$

but it follows from the method of fitting that

$$\Sigma(dx_2) = 0$$
$$\Sigma(dx_3) = 0$$
$$\Sigma(dx_4) = 0$$

and, therefore, $\Sigma(d^2) = -\Sigma(dx_1)$ $\ldots\ldots (2)$

Multiplying each residual equation (1) by x_1 and adding, we have

$$\Sigma(dx_1) = b_{12.34}\Sigma(x_1x_2) + b_{13.24}\Sigma(x_1x_3) + b_{14.23}\Sigma(x_1x_4) - \Sigma(x_1^2)$$

(Footnote continued on page 496)

Substituting the proper values, we have

$$S^2_{1.234} = 1188.688 - 100.482 - 458.700 - 159.369$$
$$= 470.137$$
$$S_{1.234} = 21.68$$

This is to be interpreted just as the standard error was interpreted in previous cases. The reliability of estimates based upon the mean value of X_1 is measured by σ_1, which has a value of 34.477. The reliability of estimates based upon the equation of net regression, when yield is considered as a function of temperature in June, July and August, is measured by $S_{1.234}$ which has a value of 21.68. It is clear that estimates made from the equation are distinctly more reliable than those based upon a knowledge of X_1 alone. We have by no means accounted for all the factors that are responsible for variability in corn yield, but we have measured and reduced to precise terms the effect of three factors upon the yield of corn per acre in Kansas.

The Coefficient of Multiple Correlation

We have need now of our third measure, the abstract coefficient of correlation. The value of this coefficient, as we have seen, depends upon the relation between S and σ.

Substituting the equivalent of $\Sigma(dx_1)$ in equation (2) we secure

$$\Sigma(d^2) = \Sigma(x_1^2) - b_{12.34}\Sigma(x_1x_2) - b_{13.24}\Sigma(x_1x_3) - b_{14.23}\Sigma(x_1x_4)$$

$$S^2_{1.234} = \frac{\Sigma(d^2)}{N} = \frac{\Sigma(x_1^2)}{N} - b_{12.34}\frac{\Sigma(x_1x_2)}{N} - b_{13.24}\frac{\Sigma(x_1x_3)}{N} - b_{14.23}\frac{\Sigma(x_1x_4)}{N}$$

Since the variables refer to deviations from the means, we have

$$S^2_{1.234} = \sigma_1^2 - b_{12.34}p_{12} - b_{13.24}p_{13} - b_{14.23}p_{14}$$

The standard error may also be derived from the equation in the general form, with the origin at zero on the original scales. For this general case we have

$$S^2_{1.234\ldots n} = \frac{\Sigma(X_1^2) - a\Sigma(X_1) - b_{12.34\ldots n}\Sigma(X_1X_2) - b_{13.24\ldots n}\Sigma(X_1X_3) - b_{14.23\ldots n}\Sigma(X_1X_4) - \ldots}{N}$$

A proof of these relations is set forth in Appendix A.

It may be computed in the present instance from the formula

$$R^2_{1.234} = 1 - \frac{S^2_{1.234}}{\sigma_1^2}$$

When the relationship between a single dependent variable and several independent variables is being studied, this measure is termed the coefficient of multiple correlation and is represented by the symbol R. The subscript to the left of the period relates to the dependent variable, while those to the right relate to the independent variables. Substituting in this formula the equivalent of $S^2_{1.234}$, we have

$$R^2_{1.234} = 1 - \frac{\sigma_1^2 - b_{12.34}p_{12} - b_{13.24}p_{13} - b_{14.23}p_{14}}{\sigma_1^2}$$

which reduces to [1]

$$R^2_{1.234} = \frac{b_{12.34}p_{12} + b_{13.24}p_{13} + b_{14.23}p_{14}}{\sigma_1^2}$$

Inserting the proper values we have

$$R^2_{1.234} = \frac{100.482 + 458.700 + 159.369}{1188.688}$$

$$R_{1.234} = .778$$

The coefficient of multiple correlation is an index of the degree of relationship between a single dependent variable and a number of independent variables, in combination. It measures the degree to which variations in the dependent variable are related to the combined action of the other factors. Its significance may be clearer if all the independent variables are looked upon as constituting a single independent series. The coefficient is then seen to be a measure of the relationship between the dependent variable and the independent series, which is precisely what the coefficient

[1] The coefficient of multiple correlation may also be derived from the general formula, which refers to an origin at zero on the original scales. This general formula is

$$R^2_{1.234\ldots n} = \frac{a\Sigma(X_1) + b_{12.34\ldots n}\Sigma(X_1X_2) + b_{13.24\ldots n}\Sigma(X_1X_3) + b_{14.23\ldots n}\Sigma(X_1X_4) + \ldots - Nc_1^2}{\Sigma(X_1)^2 - Nc_1^2}$$

of correlation is in the simpler case of two variables. In the multiple case the independent series has several component elements, but this fact does not alter the essential significance of the coefficient. No positive or negative sign is attached to R, it should be noted. In the present instance all of the independent variables are negatively correlated with corn yield, and a negative sign might be attached. It might be the case, however, that the correlation were positive for some of the independent variables, and negative for others. Because of this fact, R is always given without sign. The signs of the constants in the equation of net regression show which of the independent variables are positively correlated and which are negatively correlated with the dependent variable.

Comparison of Measures of Relationship

The degree to which our knowledge of the causes of variation in corn yield has been improved and the reliability of our estimates increased by taking account of the various factors in combination may be more readily appreciated if we bring together the various measures secured in the course of this analysis.

Table 120

A Comparison of Certain Measures Pertaining to the Corn Yield in Kansas

Basis of Estimate	*Measure of reliability of estimate*	*Coefficient of Correlation*
Arithmetic mean of X_1	$\sigma_1 = 34.477$	
$X_1 = 522.31 - 5.743X_2$	$S_{12} = 30.22$	$r_{12} = -.4814$
$X_1 = 827.64 - 9.302X_3$	$S_{13} = 24.73$	$r_{13} = -.6968$
$X_1 = 517.86 - 6.098X_4$	$S_{14} = 29.98$	$r_{14} = -.4937$
$X_1 = 1091.94 - 2.095X_2 - 7.394X_3 - 3.354X_4$	$S_{1.234} = 21.68$	$R_{1.234} = .778$

The value of S might be further reduced and the value of R correspondingly increased by bringing into the analysis other factors, such as rainfall during the growing months. The method which has been explained may be extended

to cover any number of variables, one equation being added to the set of simultaneous equations for each additional variable introduced.

The Method of Multiple Correlation Valid for Linear Relationships

One important condition has not been emphasized in the course of the preceding discussion. The validity of this method of multiple correlation depends upon the existence of a linear relationship between each pair of variables. Thus with four variables there were six pairings possible (i.e. six mean products were computed). If there had been a material departure from linearity in any of these six relationships the significance of the results would have been decreased. There would be no fallacy involved in the use of the equation under these conditions, but it would not furnish as good a basis for estimates as one which took account of the true relationship. In such a case the values of S and R would indicate that the estimates based upon the assumption of linear relationship were not very reliable.[1]

An Application of the Method

Let us illustrate the use of the estimating equation. In the year 1922 the average June temperature in Kansas was 75.2° F., the average July temperature was 77° F. and the average August temperature was 80.1° F. What was the probable corn yield per acre? Substituting these values for X_2, X_3, and X_4 in the equation,

$$X_1 = 1091.94 - 2.095X_2 - 7.394X_3 - 3.354X_4$$

[1] An approach to problems of multiple correlation when the relationship between the subordinate series is non-linear is explained by M. J. B. Ezekiel in "A Method of Handling Curvilinear Correlation for any Number of Variables." *Journal of the American Statistical Association*, Vol. XIX, N. S. No. 148, 1924.

we have

$$X_1 = 1091.94 - (2.095 \times 75.2) - (7.394 \times 77.0) - (3.354 \times 80.1)$$

$$X_1 = 96.4$$

This result expresses the actual yield as a percentage of normal. The equation to the line of trend of corn yield is

$$Y = 19.967 - .1502X$$

(where Y represents the yield and X the time factor, with origin at 1906). From this equation we secure the normal yield for 1922, finding it to be 17.6 bu. per acre. The estimated yield is 96.4 per cent of this, or 17 bu. per acre.

What are the limits within which we may expect the actual yield to fall, with respect to this estimate? The value of $S_{1,234}$ is 21.68. This means that the odds are 68 out of 100 that the actual yield ratio will be within the limits $96.4 + 21.68$, or 118.1, and $96.4 - 21.68$, or 74.7. These values refer to actual yield as percentage of normal. Knowing the normal to be 17.6, these limits may be put in terms of bushels. In this form, we have 20.8 bushels per acre and 13.1 bushels per acre as the limits within which the actual value should fall in 68 cases out of 100. The actual yield in 1922 was 19.3 bu. per acre.

In this illustration we have used one of the years included in the study. The same method would be employed in making an estimate for some future year, but in such a case an additional element of uncertainty would be introduced by the projection of the line of trend.

The Meaning of Partial or Net Correlation

In the preceding section we have sought to determine the degree to which corn yield in Kansas is affected by the temperature in June, July and August, treating the three independent variables in combination. Our aim has been to measure their combined effect upon corn yield. There

is a related problem, which in many studies may be of major importance. This is the determination of the relationship between a dependent variable and a single independent variable *when all other factors are held constant.* Concretely, what would be the effect upon corn yield of variations in July temperature, if June temperature and August temperature, and all other factors affecting yield, could be held constant? This is the problem of *net* or *partial correlation.*

It is obvious that if a method could be developed by which a single factor could be thus isolated for separate study, it would add immeasurably to the analytical powers of the economist, and of social scientists in general. It would give to the student in these fields that power to eliminate irrelevant influences and to concentrate his attention upon a single factor which is possessed by the chemist, for example. In studying the effect of one element upon another the chemist seeks to eliminate all other elements, and the effectiveness of his analysis depends in large part upon the degree to which it is possible thus to isolate the object of immediate interest.

It is not generally possible in economic analysis to eliminate all but one of the factors responsible for variations in a given series. The direct and indirect causes of a given economic phenomenon are too numerous and too complicated in their interaction for the economist ever to hope to emulate the chemist in reducing his problem to terms of but two variables. But, within certain limits, the statistician is able to employ the method of the physical scientist in holding constant certain factors while the effects of variations in another are studied. The methods which make this possible are among the most powerful of the instruments which the student of the social sciences possesses.

The method of partial correlation may be explained with reference to the problem of corn yield in Kansas. Our

object is to determine the net correlation between corn yield and the temperature in each of the three months for which the average temperature is given.

Distinction between Partial and Simple Correlation

It is important to distinguish between this problem and that faced in the ordinary measurement of relationship between two variables. We have already secured, as a description of the average relationship between corn yield and July temperature, the equation

$$X_1 = 827.64 - 9.302X_3$$

with

$$S_{13} = 24.73$$

and

$$r_{13} = -.697$$

These measures describe the relationship in question when all other factors are ignored. They are not taken account of. They are merely neglected. It is as though the chemist, in studying the reaction of one element to another, used a test tube containing various impurities, which he made no attempt to remove. The economist cannot, in general, locate and remove all the "impurities" in his problem, but he should recognize that his measures relate to such uncorrected data.

The Method of Partial Correlation

In seeking to determine the net correlation between corn yield and July temperature we attempt to secure a measure of the correlation which would prevail if all other factors might be held constant. We shall take full account of the other factors we have studied, but we shall try to secure a measure which will be influenced only by fluctuations in July temperature, in relation to corn yield.

One possible method of accomplishing this end may be suggested. If one possessed data covering a very long period we might be able to pick out a number of years during which the average temperatures in June and August remained unchanged. Let us say that we could find thirty years in all during each of which the June temperature averaged 74° and the August temperature 78°. Corn yield and July temperature varied during these years. The relationship between July temperature and corn yield might now be measured, and it would be certain that the results would not be affected by the presence of fluctuations in June temperature and August temperature. Unfortunately, this method of holding certain factors constant cannot be employed. The data are too limited and too varied, in general, to enable us to pick from among them such figures as are appropriate to our purpose. Other methods of arriving at the same end are available, however.

The Relation between the Coefficients of Regression and the Coefficient of Correlation

In introducing this method, let us return for a moment to the problem of simple correlation. We have seen that the equation which describes the relationship between two variables may be expressed in two forms, one with y dependent, the other with x dependent. These equations, with variables relating to deviations from the respective arithmetic means, are of the form

$$y = r\frac{\sigma_y}{\sigma_x}x$$

and

$$x = r\frac{\sigma_x}{\sigma_y}y$$

In these equations the values $r\frac{\sigma_y}{\sigma_x}$ and $r\frac{\sigma_x}{\sigma_y}$ are termed the

coefficients of regression. They measure the slopes of the two lines of regression. The coefficient of correlation is equal to the square root of these two coefficients of regression. That is

$$r = \sqrt{r\frac{\sigma_y}{\sigma_x} \cdot r\frac{\sigma_x}{\sigma_y}}.$$

Using the symbols ordinarily employed for these values we have

$$r_{yx} = \sqrt{b_{yx} \cdot b_{xy}}$$

where b_{yx} is the coefficient of regression of y on x (y dependent) and b_{xy} is the coefficient of regression of x on y (x dependent). If we use the symbols x_1 and x_2 instead of y and x, and make corresponding changes in the subscripts, we have

$$r_{12} = \sqrt{b_{12} \cdot b_{21}}$$

The value of r, it may be noted, is the same whether x_1 or x_2 be dependent. That is, $r_{12} = r_{21}$.

This leads us at once to the solution of our present problem. We seek a coefficient of net correlation between two variables, a measure of their relationship when other factors are held constant. But in the study of multiple correlation we have already derived certain coefficients of net regression. The equation of regression, which describes the relationship between a single dependent and a number of independent variables, is of the form

$$X_1 = a + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$$

The quantities $b_{12.34}$, $b_{13.24}$ and $b_{14.23}$ are the coefficients of net regression. They indicate the weight to be attached to each of the independent variables when an estimate is based upon the three in combination. Thus, the coefficient $b_{13.24}$ is a measure of the weight to be attached to the variable X_3 in estimating X_1, when variables X_2 and X_4 are also taken account of. We may secure the values of these three coefficients, as we have seen, by the simple method of solving simultaneous equations.

To get the desired coefficients of net correlation by this method, however, we need additional values. The coefficient r_{yx} is the square root of b_{yx} and b_{xy}. The coefficient r_{13} is the square root of b_{13} and b_{31}. Similarly, the coefficient we desire, which measures the degree of relationship between X_1 (corn yield) and X_3 (July temperature) when X_2 (June temperature) and X_4 (August temperature) are held constant, may be derived from the formula:

$$r_{13.24} = \sqrt{b_{13.24} \cdot b_{31.24}}$$

In this equation $b_{13.24}$ is the net regression coefficient of X_1 on X_3, and $b_{31.24}$ is the net regression coefficient of X_3 on X_1.

The meaning of these symbols has been made clear in the course of the discussion. The subscripts indicate the particular relationship to which the values apply. The two subscripts to the left of the point (the primary subscripts) refer to the variables to which the measure applies specifically, while the subscripts to the right of the point (the secondary subscripts) indicate the variables which are held constant for the purpose of the particular comparison being made. The number held constant is two, in the present case, though it might be one, or any other number. Thus the general formula for the coefficient of net correlation between variables X_1 and X_3 would be written

$$r_{13.2456 \cdots n} = \sqrt{b_{13.2456 \cdots n} \cdot b_{31.2456 \cdots n}}$$

Of the two primary subscripts the first refers to the dependent variable, the second to the independent variable.

To solve for $r_{13.24}$ in the problem before us only one additional value is needed. We have the value of the coefficient $b_{13.24}$, having obtained this in the preceding illustration. The coefficient of regression $b_{31.24}$ is one of the constants in the equation:

$$X_3 = a + b_{31.24}X_1 + b_{32.14}X_2 + b_{34.21}X_4$$

which is the equation describing the average relationship between July temperature, as dependent, and June temperature, August temperature and corn yield as independent variables. Given certain values of X_1, X_2 and X_4, this equation enables us to estimate the most probable July temperature which would accompany these values. (It describes an association, rather than a relation of dependence, when expressed in this way.)

Solution of Normal Equations

The required values are secured just as in the example presented above. Four normal equations are secured in the usual fashion. These may be reduced to the three which follow:

$$p_{13} = \sigma_1^2 b_{31.24} + p_{12} b_{32.14} + p_{14} b_{34.21}$$
$$p_{23} = p_{12} b_{31.24} + \sigma_2^2 b_{32.14} + p_{24} b_{34.21}$$
$$p_{34} = p_{14} b_{31.24} + p_{24} b_{32.14} + \sigma_4^2 b_{34.21}$$

The mean products and the standard derivations needed for insertion in these equations have all been used in the solution of the preceding set, with the single exception of σ_1^2. Inserting these values we have

$$-62.039 = 1188.688 b_{31.24} - 47.967 b_{32.14} - 47.519 b_{34.21}$$
$$2.790 = -47.967 b_{31.24} + 8.3564 b_{32.14} + 2.932 b_{34.21}$$
$$2.063 = -47.519 b_{31.24} + 2.932 b_{32.14} + 7.7893 b_{34.21}$$

Solving these equations we secure the following values:

$$b_{31.24} = -.0531$$
$$b_{32.14} = +.0574$$
$$b_{34.21} = -.0807$$

Computation of Coefficient of Net Correlation

From the preceding example we have the value

$$b_{13.24} = -7.394$$

Substituting in the formula

$$r_{13.24} = \sqrt{b_{13.24} \times b_{31.24}}$$

we have

$$r_{13.24} = \sqrt{-7.394 \times -.0531}$$
$$= -.6266$$

(As in the simpler case in which only two variables are dealt with, r takes the sign of the coefficients of regression, which will be alike in this respect. If both regression coefficients are positive, r is positive, if both are negative, r is negative.)

This is the coefficient of net correlation between corn yield and July temperature in Kansas. It measures the degree of relationship between these two variables when June temperature and August temperature are held constant.

Another Method of Computing Coefficients of Partial Correlation

Obviously a whole series of such coefficients can be worked out in dealing with a number of variables. In computing a number of such measures a method may be utilized which differs somewhat from that employed above, and which has certain advantages in the way of systematic arrangement.

A simple coefficient of correlation relating to but two variables is termed a *coefficient of zero order*. Such coefficients are represented by symbols of the type r_{12}, r_{24}, etc. Coefficients of net correlation which relate to two variables, while a single additional variable is held constant, are termed *coefficients of the first order*, and are represented by symbols such as $r_{12.3}$, $r_{24.3}$, etc. Similarly, we may have coefficients of the second, third, fourth or nth order, depending upon the number of variables held constant while the relationship between a single dependent and a single independent variable is being measured.

It is possible to derive each coefficient of partial correla-

tion from those of the next lower order. Thus a coefficient of the first order may be derived from the relation

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{(1 - r^2_{13})^{\frac{1}{2}} (1 - r^2_{23})^{\frac{1}{2}}}$$

For a coefficient of the second order

$$r_{12.34} = \frac{r_{12.3} - r_{14.3} \cdot r_{24.3}}{(1 - r^2_{14.3})^{\frac{1}{2}} (1 - r^2_{24.3})^{\frac{1}{2}}}$$

As a general equation for a coefficient of net correlation of any order, we have

$$r_{12.345 \cdots n} = \frac{r_{12.345 \cdots (n-1)} - r_{1n.345 \cdots (n-1)} \cdot r_{2n.345 \cdots (n-1)}}{(1 - r^2_{1n.345 \cdots (n-1)})^{\frac{1}{2}} (1 - r^2_{2n.345 \cdots (n-1)})^{\frac{1}{2}}}$$

Thus it is possible, starting with the zero order coefficients of correlation, to compute all higher order coefficients successively. The mere arithmetic of calculation would be laborious, but certain prepared tables reduce these computations to a minimum.[1] The method may be illustrated, using the data of the preceding problem.

In the present case we require three coefficients of the second order, $r_{12.34}$, $r_{13.24}$ and $r_{14.23}$. These will serve as measures of the net correlation between corn yield and temperature in each of the three critical months. The formula from which the first of these measures may be computed was given above. For the second, we have

$$r_{13.24} = \frac{r_{13.2} - r_{14.2} \cdot r_{34.2}}{(1 - r^2_{14.2})^{\frac{1}{2}} (1 - r^2_{34.2})^{\frac{1}{2}}}$$

and for the third

$$r_{14.23} = \frac{r_{14.2} - r_{13.2} \cdot r_{43.2}}{(1 - r^2_{13.2})^{\frac{1}{2}} (1 - r^2_{43.2})^{\frac{1}{2}}}$$

But each of these values may be derived from a slightly different grouping of first order coefficients. We may use the three formulas

[1] J. R. Miner, *Tables of* $\sqrt{1 - r^2}$ *and* $1 - r^2$ *for use in Partial Correlation and in Trigonometry*, Johns Hopkins Press, Baltimore, Md., 1922.

$$r_{12.34} = \frac{r_{12.4} - r_{13.4} \cdot r_{23.4}}{(1 - r^2_{13.4})^{\frac{1}{2}} (1 - r^2_{23.4})^{\frac{1}{2}}}$$

$$r_{13.24} = \frac{r_{13.4} - r_{12.4} \cdot r_{32.4}}{(1 - r^2_{12.4})^{\frac{1}{2}} (1 - r^2_{32.4})^{\frac{1}{2}}}$$

$$r_{14.23} = \frac{r_{14.3} - r_{12.3} \cdot r_{42.3}}{(1 - r^2_{12.3})^{\frac{1}{2}} (1 - r^2_{42.3})^{\frac{1}{2}}}$$

By employing both methods in computing each second order coefficient a check upon the calculations is afforded.

Computation of First Order Coefficients

The second order coefficients cannot be computed until all necessary first order coefficients have been secured. The necessary equations, of the type

$$r_{12.3} = \frac{r_{12} - r_{13} \cdot r_{23}}{(1 - r^2_{13})^{\frac{1}{2}} (1 - r^2_{23})^{\frac{1}{2}}},$$

may be constructed from the general formula for coefficients of partial correlation. Since several of these values must be computed, a systematic procedure should be adopted. The tabular arrangement on page 510 facilitates the computations and lends itself readily to the employment of the tables mentioned above.

The procedure in computing each first order coefficient is simple. Three zero order coefficients are necessary for each calculation. These should be arranged in the table in the order in which they occur in the numerator of the fraction from which the required coefficient is to be computed. The numerator of this fraction is secured by subtracting from the first zero order coefficient the product of the other two. This product term appears in one column of the table. The denominator of the fraction is the product of two terms of the type $\sqrt{1 - r^2}$, derived from the second and third coefficients in each group of three. The tabular arrangement permits these computations to be carried forward systematically.

TABLE 121

Illustrating the Computation of First Order Coefficients of Partial Correlation

Kansas Corn Yield and Temperature

r 0 *Order*		$(1-r^2)^{\frac{1}{2}}$	*Product term of numerator*	*Whole numerator*	*Denominator*	*r 1st Order*	
Subscript	*Coefficient*					*Subscript*	*Coefficient*
12	−.4814		−.2604	−.2210	.6653	12.3	−.3322
13	−.6968	.7173					
23	+.3737	.9275					
14	−.4937		−.1994	−.2943	.6873	14.3	−.4282
13	−.6968	.7173					
43	+.2862	.9582					
24	+.3633		+.1070	+.2563	.8887	24.3	+.2884
23	+.3737	.9275					
43	+.2862	.9582					
13	−.6968		−.1799	−.5169	.8130	13.2	−.6358
12	−.4814	.8765					
32	+.3737	.9275					
14	−.4937		−.1749	−.3188	.8166	14.2	−.3904
12	−.4814	.8765					
42	+.3633	.9317					
34	+.2862		+.1358	+.1504	.8642	34.2	+.1740
32	+.3737	.9275					
42	+.3633	.9317					
12	−.4814		−.1794	−.3020	.8102	12.4	−.3727
14	−.4937	.8696					
24	+.3633	.9317					
13	−.6968		−.1413	−.5555	.8333	13.4	−.6666
14	−.4937	.8696					
34	+.2862	.9582					
23	+.3737		+.1040	+.2697	.8928	23.4	+.3021
24	+.3633	.9317					
34	+.2862	.9582					
14	−.4937		−.1994	−.2943	.6873	14.3	−.4282
13	−.6968	.7173					
43	+.2862	.9582					
12	−.4814		−.2604	−.2210	.6653	12.3	−.3322
13	−.6968	.7173					
23	+.3737	.9275					
42	+.3633		+.1070	+.2563	.8887	42.3	+.2884
43	+.2862	.9582					
23	+.3737	.9275					

The coefficient $r_{23.4}$ is, of course, identical with $r_{32.4}$; $r_{34.2}$ is identical with $r_{43.2}$, etc. It is unnecessary to duplicate the work of computation with respect to these measures.

Computation of Second Order Coefficients

From these first order coefficients the three required second order coefficients may be secured by methods analogous to those employed above. The computations are shown in the following table. As a check upon the calculations each required measure is computed from two different combinations of the first order coefficients.

Table 122

Illustrating the Computation of Second Order Coefficients of Partial Correlation

Kansas Corn Yield and Temperature

r 1st Order		$(1-r^2)^{\frac{1}{2}}$	*Product term of numerator*	*Whole numerator*	*Denominator*	*r* 2nd Order	
Subscript	*Coefficient*					*Subscript*	*Coefficient*
12.3	−.3322		−.1235	−.2087	.8653	12.34	−.2412
14.3	−.4282	.9037					
24.3	+.2884	.9575					
13.2	−.6358		−.0679	−.5679	.9065	13.24	−.6265
14.2	−.3904	.9206					
34.2	+.1740	.9847					
14.2	−.3904		−.1106	−.2798	.7601	14.23	−.3681
13.2	−.6358	.7719					
43.2	+.1740	.9847					
12.4	−.3727		−.2014	−.1713	.7106	12.34	−.2411
13.4	−.6666	.7454					
23.4	+.3021	.9533					
13.4	−.6666		−.1126	−.5540	.8847	13.24	−.6262
12.4	−.3727	.9280					
32.4	+.3021	.9533					
14.3	−.4282		−.0958	−.3324	.9031	14.23	−.3681
12.3	−.3322	.9432					
42.3	+.2884	.9575					

The value of $r_{13.24}$, it will be noted, is the same as that derived from the two coefficients of net regression.

The meaning of such coefficients as these was explained in the earlier section dealing with this problem. The following summary of results may make clear the gain in knowledge which has resulted from the above analysis.

$r_{12} = -.4814$	$r_{12.34} = -.2412$
$r_{13} = -.6968$	$r_{13.24} = -.6265$
$r_{14} = -.4937$	$r_{14.23} = -.3681$

It is clear that the net effect of June temperature upon corn yield is distinctly less than was indicated by the simple correlation. This is so because there is a positive correlation between temperature in June and temperature in July and August, so that the crude correlation of two variables alone shows June temperature as more important than it really is. For the same reason, all the net coefficients are less than the simple coefficients, though it is still apparent that July temperature is far more important, in relation to corn yield, than the temperature in either of the other months.

The coefficients of net correlation are net, of course, only with respect to the variables actually taken account of, and held constant. Thus there may be other factors, such as rainfall in June, July or August, which affect corn yield and which are correlated with the temperature during these months. Were these included the various coefficients of net correlation might have higher or lower values than those given.

A Measure of Variability

Having these coefficients of net correlation, another measure of some importance may be computed. This is a measure of the variability of a single character while a number of related variables are held constant. Thus the question might arise: If we could hold constant the tempera-

ture in Kansas in June, July and August, what would be the variability of the corn yield? In other words: If we could eliminate such variability in corn yield as is due to variability in temperature, what fluctuations would remain in the yield of corn? This measure of variability is represented by the symbol $\sigma_{1.234\ldots n}$. It is termed the standard deviation of order n.

This measure may be computed from the general equation

$$\sigma^2_{1.23\ldots n} = \sigma_1^2(1 - r^2_{12})\,(1 - r^2_{13.2})\,(1 - r^2_{14.23})\ldots(1 - r^2_{1n.23\ldots n-1})$$

Applying this formula to the results of the study of corn yield, we have

$$\sigma^2_{1.234} = 1188.688[1 - (-.4814)^2][1 - (-.6358)^2][1 - (-.3681)^2]$$

$$\sigma^2_{1.234} = 470.3364$$

$$\sigma_{1.234} = 21.69$$

Referring back to the discussion of this problem we find that the values of $\sigma_{1.234}$ and $S_{1.234}$ are identical. (There is a difference of .01, as derived above.) That is, the standard deviation of variable X_1, when variables X_2, X_3 and X_4 are held constant, is merely the measure of scatter about the line of regression. It is the standard error of estimate, when estimates are based upon the factors X_2, X_3 and X_4. The reason for this is obvious. The variability of the original series is reduced to the extent that estimates based upon the equation of relationship approximate the actual values. The variability which remains is due to differences between these estimates and the actual values. But these differences are merely the deviations from the line of regression, from which S is computed. A realization of the identity of these two measures may assist in making their meaning clear.

Since $\sigma_{1.234}$ and $S_{1.234}$ are identical, the coefficient of multiple correlation, $R_{1.234}$, may be computed from the equation

$$R^2_{1.23\ldots n} = 1 - \frac{\sigma^2_{1.23\ldots n}}{\sigma_1^2}$$

or, using the formula for $\sigma^2_{1.234\ldots n}$, from the equation

$$1 - R^2_{1.23\ldots n} = (1 - r^2_{12})\,(1 - r^2_{13.2})\,(1 - r^2_{14.23})\ldots (1 - r^2_{1n.23\ldots(n-1)})$$

The method first presented, based directly upon the method of least squares, furnishes a simpler approach to the problem, however, and is in general to be preferred.

Certain Limitations

The measures we have described in dealing with problems of multiple and partial correlation are valid on the assumption that the relationship between the different variables is in all cases linear. Thus with four variables six different pairs may be obtained. The regression in each of these six cases should be linear if combined or net effects are to be studied by the methods outlined above. If the regression is non-linear when natural numbers are dealt with, it may be possible to secure linear relationships by correlating logarithms or reciprocals. Thus we might derive an estimating equation of the type

$$\text{Log } X_1 = a + b_{12.34}X_2 + b_{13.24}X_3 + b_{14.23}X_4$$

if the relation between X_1 in logarithmic form and each of the other variables in the original arithmetic form were linear. The corresponding measures, S and R, would then relate to ratios, as in the examples given in the preceding chapter.

One other important limitation should be noted. Coefficients of multiple or of net correlation based upon a large number of variables have little significance unless the number of observations be large. Misleadingly high values will be secured when studies involving many variables are based upon small samples. Within the limits set by these restrictions, the methods of multiple and partial correlation constitute very powerful instruments for economic analysis.

REFERENCES

BOWLEY, ARTHUR L. *Elements of Statistics* (398–408).

EDGEWORTH, F. Y. *On Correlated Averages.* Phil. Mag. 5th series, Vol. XXXIV, 1892 (194).

EZEKIEL, M. J. B. *A Method of Handling Curvilinear Correlation for any Number of Variables.* Journal of the American Statistical Association, Vol. XIX, N. S. 148, 1924.

HAAS, G. C. *Sale Prices as a Basis for Farm Land Appraisal.* Technical Bulletin No. 9. University of Minnesota Agricultural Experiment Station.

KELLEY, TRUMAN L. *Statistical Method* (279–310).

KELLEY, TRUMAN L. *Partial and Multiple Correlation* (in Rietz, H. L. *Handbook of Mathematical Statistics,* 139–149).

MINER, J. R. *Tables of* $\sqrt{1-r^2}$ *and* $1-r^2$ *for use in Partial Correlation and Trigonometry.*

PEARL, RAYMOND. *Medical Biometry and Statistics* (319–331).

PEARSON, KARL. *Regression, Heredity and Panmixia.* Phil. Transactions Royal Society, Series A. Vol. CLXXXVII, 1896 (253–318).

SMITH, BRADFORD B. *The Use of Punched Card Tabulating Equipment in Multiple Correlation Problems.* (Prepared for the use of statisticians of the Bureau of Agricultural Economics, U. S. Dept. of Agriculture.)

TOLLEY, H. R. and EZEKIEL, M. J. B. *A Method of Handling Multiple Correlation Problems.* Journal of the American Statistical Association, Dec. 1923.

YULE, G. U. *An Introduction to the Theory of Statistics* (229–253). (The notation usually employed in the correlation of several variables was developed by Yule. It is explained in this reference.)

CHAPTER XV

ELEMENTARY PROBABILITIES AND THE NORMAL CURVE OF ERROR

Reference has been made in an earlier section to the family resemblance which is found in frequency distributions drawn from widely different fields. Attention was also drawn to a certain basic type which is represented graphically by the symmetrical bell-shaped curve which is called the "normal curve," or the "normal curve of error." In an earlier day this curve was looked upon as representing a fundamental law which described all distributions of quantitative data. From the modern standpoint this was quite an erroneous conception. The normal curve is viewed today as but one of a number of types of curves which may be used to describe frequency distributions. It is, however, by far the most important type, and an understanding of its characteristics is essential to the statistician.

Elementary Theorems in Probability

This understanding may best be secured by a brief consideration of certain elementary principles of probability. A detailed explanation of the theory of probability would carry us far beyond the limits of the present volume. The treatment which follows is presented only as an introduction to the subject, designed to illustrate, by simple numerical examples, the relation between the principles of probability and the normal law of error.

In this argument we may use the following standard notation. If an event can occur in n ways, a of which are to be considered as successful and b as unsuccessful, the probability p of a successful outcome may be written

$$p = \frac{a}{n}$$

and the probability q of an unsuccessful outcome may be written.

$$q = \frac{b}{n}$$

Since the sum of the favorable and unfavorable outcomes is equal to the total number of events, we have

$$a + b = n$$

Dividing by n,

$$\frac{a}{n} + \frac{b}{n} = 1$$

so that

$$p + q = 1$$

or *certainty*.

A probability, therefore, may be written as a ratio. The numerator of the fraction corresponding to this ratio represents the number of favorable (or unfavorable) outcomes, while the denominator represents the total number of possible outcomes.

Examples of Simple Probabilities

If a coin be tossed, the turning up of a head being looked upon as a favorable outcome, we have, as the probability of a success,

$$p = \frac{1}{2}$$

and of a failure,

$$q = \frac{1}{2}$$

If we roll a die, regarding a six spot as a favorable outcome,

$$p = \frac{1}{6}$$

and

$$q = \frac{5}{6}$$

If a card be drawn from a pack of 52 the chance of drawing the ace of spades is $\frac{1}{52}$, of failing in that endeavor, $\frac{51}{52}$.

The Addition of Probabilities

What is the chance of securing *either* an ace of spades *or* a two of spades in a single draw from a pack of 52 cards? In such a case, where any one of several outcomes will be considered as favorable, the probability of a success is the sum of the separate probabilities. In this example

$$p = \frac{1}{52} + \frac{1}{52} = \frac{1}{26}$$

The chance of drawing either a heart or a spade from a pack of playing cards is given by

$$p = \frac{13}{52} + \frac{13}{52} = \frac{1}{2}$$

The Multiplication of Probabilities

Two events are said to be independent when the outcome of one does not affect the outcome of the other. Thus the result of one throw of a die does not, presumably, affect the result of the next toss. The probability of a *compound event* (i.e. that two events, independent of one another, will both occur) is the *product* of the probabilities of the separate events. Thus the chance of securing an ace, followed by a two spot, in two successive throws of a die, is given by

$$p = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$$

In computing the probability of a given outcome, it is frequently necessary both to multiply and to add probabili-

ties. For example, we wish to determine the chance of securing the total 5 from two dice thrown simultaneously. We may label the dice *a* and *b* to distinguish them. This total may be secured from any one of the four following combinations

Die *a*	Die *b*
1	4
2	3
3	2
4	1

The chance of securing an ace with die *a* is $\frac{1}{6}$, of securing a 4 with die *b* is $\frac{1}{6}$. The chance of the two in combination is $\frac{1}{36}$. Similarly, the probability of each of the other three combinations is $\frac{1}{36}$. But any one of these four results will give a total of 5, and will be considered successful. Hence

$$p = \frac{1}{36} + \frac{1}{36} + \frac{1}{36} + \frac{1}{36} = \frac{1}{9}$$

We have in this example answered the question: What is the probability of securing exactly 5 in the toss of two dice? We might put the question: What is the chance of securing *at least* 5 in the toss of two dice? In this case a total of 5 or more will be considered a favorable outcome. Just as in the preceding example, we may work out the probability of securing each of the results which will be accepted as successful. The following summary indicates the probability of each of these totals:

Probability of throwing 12 with two dice $= \frac{1}{36}$

" " " 11 " " " $= \frac{2}{36}$

" " " 10 " " " $= \frac{3}{36}$

" " " 9 " " " $= \frac{4}{36}$

Brought forward from preceding page $\frac{10}{36}$

Probability of throwing 8 with two dice $= \frac{5}{36}$

" " " 7 " " " $= \frac{6}{36}$

" " " 6 " " " $= \frac{5}{36}$

" " " 5 " " " $= \frac{4}{36}$

Sum of above probabilities $= \frac{30}{36}$

The chance of throwing at least 5 in the toss of two dice is, therefore, $\frac{30}{36}$ or $\frac{5}{6}$.

The Binomial Expansion and the Measurement of Probabilities

It is possible to express these facts in a generalized form. A simple illustration may be employed to exemplify the derivation of the general expression.

If two coins are tossed simultaneously there are four possible outcomes

a	*b*	*a*	*b*	*a*	*b*	*a*	*b*
T	*T*	*T*	*H*	*H*	*T*	*H*	*H*.

(The two coins are represented, respectively, by the letters *a* and *b*.) The chances of securing no heads, one head and two heads are, respectively, $\frac{1}{4}$, $\frac{1}{2}$ and $\frac{1}{4}$. If three coins (represented by the letters *a*, *b* and *c*) are tossed simultaneously, we have eight possible outcomes

a b c	*a b c*	*a b c*	*a b c*	*a b c*	*a b c*	*a b c*	*a b c*
TTT	*TTH*	*THH*	*THT*	*HTT*	*HTH*	*HHT*	*HHH*.

The chances of securing no heads, 1 head, 2 heads and 3 heads are, respectively, $\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}$.

But these results may be derived without the necessity of working out the separate probabilities in detail. We have employed p and q to represent, respectively, the probability of success and failure of a given event. If there are two independent events, the probabilities being the same in each case, the compound probabilities are given by the expansion of the expression

$$(p + q)^2$$

In the present case p (the probability of throwing a head) $= q = \frac{1}{2}$, and the probabilities of the various results are given by

$$\left(\frac{1}{2} + \frac{1}{2}\right)^2 = \frac{1}{4} + \frac{1}{2} + \frac{1}{4}$$

These are the results secured in the first example. If there are three independent events we have

$$\left(\frac{1}{2} + \frac{1}{2}\right)^3 = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} + \frac{1}{8},$$

the probabilities secured in the second example.

If we wish to know not the separate probabilities but the probable frequencies of the various outcomes in a given number of trials, these may be computed from the expression

$$N(p + q)^n$$

where N represents the number of trials and n the number of independent events. Thus if there are 200 trials and there are two independent events, the probable frequencies are given by

$$200(p + q)^2 = 200(p^2 + 2pq + q^2).$$

With $p = q = \frac{1}{2}$ this gives us

$$200\left(\frac{1}{4}\right) + 200\left(\frac{1}{2}\right) + 200\left(\frac{1}{4}\right) = 50 + 100 + 50$$

which indicates the probable frequencies of 2 successes, 1 success and no successes.

If there are three independent events, the probable frequencies in N trials are determined from the binomial expansion of

$$N(p+q)^3$$

If N equals 200, we have

$$200(p^3+3p^2q+3pq^2+q^3)$$

If p equals $\frac{1}{2}$, we have

$$200\left(\frac{1}{8}\right)+200\left(\frac{3}{8}\right)+200\left(\frac{3}{8}\right)+200\left(\frac{1}{8}\right)$$

These terms indicate, in order, the probable frequencies of 3 successes, 2 successes, 1 success and no successes. The total frequencies secured by carrying through the process of multiplication will be equal to the number of trials, for all possible outcomes are covered by the expansion.

Thus, when we know the probabilities in advance,[1] we may determine the probable frequencies of any given number of successes or failures, a fact of great significance in the development of statistical theory.

A Comparison of Actual and Theoretical Frequencies in the Realm of Pure Chance

Certain points of importance may be made clear by comparing some experimental results with the theoretical

[1] A distinction is generally drawn between *a priori* probabilities of the type described above, and *empirical* probabilities, knowledge of which is derived from observation or experience. As an example of the latter type we have, as the probability that a man aged 35 will live 10 years, the ratio $\frac{74173}{81822}$. This is based upon the American Experience Table of Mortality which shows that of 81,822 men living at the age of 35, there are 74,173 living ten years later. Chapter XVI is devoted to certain considerations relating to the use of such empirical measures.

frequencies given by the binomial expansion. Twelve dice were thrown a number of times. Each 4, 5 or 6 spot appearing was considered to be a success while a 1, 2 or 3 spot was a failure. (In a typical throw we might have the following spots up: 3, 1, 5, 1, 2, 4, 4, 6, 3, 2, 3, 5. In this lot there are five successes, and the result is so tallied.) In a classical example recorded by W. F. R. Weldon[1] twelve dice were thrown in this way 4096 times, a success being defined as above. The results are recorded in column (2) of Table 123, and the distribution is shown in Fig. 85. By computation we find the arithmetic mean and the standard deviation of this distribution to be, respectively, 6.139 and 1.712.

Let us compare with these results those which we might expect from the given conditions. Twelve dice were thrown each time, hence we are dealing with 12 independent events. There were 4096 trials. Since either a 4, 5 or 6 is considered a success, $p = q = \frac{1}{2}$.

For the terms in the binomial expansion we have

$$(a+b)^n = a^n + na^{n-1}b + \frac{n(n-1)}{1.2}a^{n-2}b^2 + \frac{n(n-1)(n-2)}{1.2.3}a^{n-3}b^3 + \ldots$$

In the present case we have

$$4096\left(\frac{1}{2}+\frac{1}{2}\right)^{12}$$

Expanding

$$4096\left(\frac{1}{4096}+\frac{12}{4096}+\frac{66}{4096}+\frac{220}{4096}+\frac{495}{4096}+\frac{792}{4096}+\frac{924}{4096}+\frac{792}{4096}+\frac{495}{4096}+\frac{220}{4096}+\frac{66}{4096}+\frac{12}{4096}+\frac{1}{4096}\right)$$

Completing the indicated multiplication we have the theoretical frequencies of the various possible successes in

[1] Cited by G. U. Yule, *An Introduction to the Theory of Statistics*, 5th ed., 258, and by F. Y. Edgeworth, Encycl. Brit., 11th ed., Vol. XXII, 394.

4096 throws of twelve dice. These are shown in column (3) of the following table.

TABLE 123

Comparison of Actual and Theoretical Frequencies in Dice-Rolling Experiment

(1) *Number of Successes*	(2) *Observed Frequencies*	(3) *Theoretical Frequencies*
0	0	1
1	7	12
2	60	66
3	198	220
4	430	495
5	731	792
6	948	924
7	847	792
8	536	495
9	257	220
10	71	66
11	11	12
12	0	1
	4096	4096

The distribution of the theoretical frequencies is shown in Fig. 85.

The advantages to be derived from being able to secure a standard type of distribution with which to compare the actual distributions found by experiment are many. We are able, in the first place, to determine the degree to which differences in the frequencies are due to chance fluctuations of sampling (or to a bias in the dice). We can determine what the true distribution would be if the test were continued until the results of an infinite number of throws were recorded. The results of the experiment, standing alone, are limited in their significance to the particular throws recorded. The theoretical frequencies have no such limitations, applying generally where the same initial probabilities apply. This fact, that a knowledge of the

theoretical frequencies affords a basis for *generalization* such as is not possible from the results of experiment alone, is the central and most significant of these advantages.

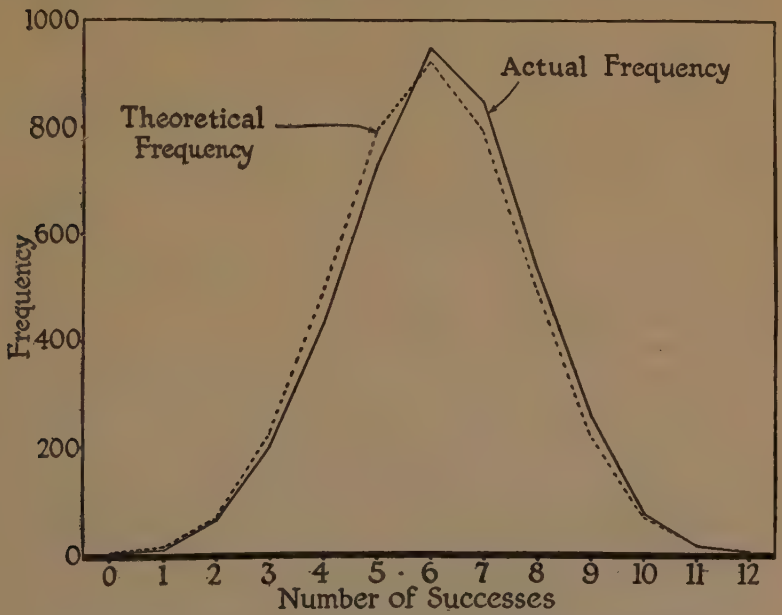


FIG. 85.—A Comparison of Actual and Theoretical Frequencies in a Dice-rolling Experiment

When we have, as in this case, a knowledge of the probabilities involved, it is possible to determine the arithmetic mean and the standard deviation of the distribution which will be formed by the theoretical frequencies. As a general expression for the mean number of successes, where the number of independent events and the probability of success are known, we have

$$M = np$$

Applying the present values,

$$M = 12 \times \frac{1}{2} = 6$$

The mean, as computed from the observed frequencies, is 6.139.

As a general expression for the standard deviation, under the same conditions, we have

$$\sigma = \sqrt{npq}$$

In the present case

$$\sigma = \sqrt{12 \times \frac{1}{2} \times \frac{1}{2}} = \sqrt{3}$$
$$= 1.732$$

The standard deviation, as computed from the actual frequencies, is 1.712.

The Normal Curve of Error

We may return to a consideration of the curve in Fig. 85 which represents the theoretical frequencies in the dice throwing experiments. It is a perfectly symmetrical 12-sided polygon, the number of sides (excluding the base) corresponding to the number of independent events in the particular problem considered. With six events we should have a six-sided figure, with twenty events a twenty-sided figure, and so on. It is obvious that, as n increases, the number of sides to the polygon increasing correspondingly in number, the graph representing the expansion of the binomial $(p + q)^n$ approaches more and more closely a smooth curve. With n infinitely large a perfectly smooth curve would be secured. This is the normal curve of error which has been plotted in Fig 87.

A great part of the power which modern statistical technique possesses is derived from the detailed knowledge of the characteristics of this normal or Gaussian curve. From prepared tables of the integrals of this curve theoretical frequencies may be determined much more readily than by the laborious method based upon the binomial

expansion. Several examples demonstrating the practical utility of such tables are presented below.

The equation to this curve is written in several forms, of which

$$y = y_o e^{-\frac{x^2}{2\sigma^2}}$$

is a common type. In this equation y_o, the maximum ordinate, is a constant; e is a constant (the base of the Napierian logarithms) having a value of 2.7182818, while σ represents the standard deviation.[1] The maximum ordinate may be derived from the relation

$$y_o = \frac{N}{\sigma\sqrt{2\pi}}$$

hence the equation to the normal curve may be written

$$y = \frac{N}{\sigma\sqrt{2\pi}} e^{-\frac{x^2}{2\sigma^2}}$$

The uses of this curve, and of the tables of the probability integral based upon the integration of this curve, are far too varied to be enumerated at length here. A simple example may serve to introduce the subject.

An Economic Application

The statistical division of the American Telephone and Telegraph Company has made a study of the annual message use of four-party line residence message rate subscribers in Buffalo. The annual messages for each of 995 subscribers were tabulated and classified.[2] The results, together with certain computations, appear in the following table.

[1] Gauss' deduction of the error equation may be found in all standard works on the theory of least squares. Cf. references at end of Appendix A.

[2] "Introduction to Frequency Curves and Averages." *Statistical Bulletin, Statistical Methods Series, No.* 1. Issued by chief statistician, American Telephone and Telegraph Co.

TABLE 124

Annual Message Use of 995 Telephone Subscribers

Illustrating the Computation of the Moments of a Frequency Distribution

(1) *Interval of message use*[1]	(2) *Mid-point*	(3) *Frequency*	(4) *Deviation from arbitrary origin in class-interval units*	(5)	(6)	(7)	(8)
	m	f	x'	fx'	$f(x')^2$	$f(x')^3$	$f(x')^4$
0– 50	25	0	−10	0	0	0	0
50– 100	75	1	−9	−9	81	−729	6561
100– 150	125	9	−8	−72	576	−4608	36864
150– 200	175	19	−7	−133	931	−6517	45619
200– 250	225	38	−6	−228	1368	−8208	49248
250– 300	275	50	−5	−250	1250	−6250	31250
300– 350	325	95	−4	−380	1520	−6080	24320
350– 400	375	85	−3	−255	765	−2295	6885
400– 450	425	115	−2	−230	460	−920	1840
450– 500	475	132	−1	−132	132	−132	132
500– 550	525	144	0	0	0	0	0
550– 600	575	116	1	116	116	116	116
600– 650	625	79	2	158	316	632	1264
650– 700	675	54	3	162	486	1458	4374
700– 750	725	31	4	124	496	1984	7936
750– 800	775	11	5	55	275	1375	6875
800– 850	825	5	6	30	180	1080	6480
850– 900	875	6	7	42	294	2058	14406
900– 950	925	2	8	16	128	1024	8192
950–1000	975	1	9	9	81	729	6561
1000–1050	1025	1	10	10	100	1000	10000
1050–1100	1075	1	11	11	121	1331	14641
		995		−956	9676	−22952	283564

[1] As here classified an item having a value of 50 was put in the class having 50 as an upper limit. Items falling on other class limits were similarly disposed of.

THE MOMENTS OF A FREQUENCY DISTRIBUTION

Certain terms and symbols which have not been employed heretofore may be introduced at this point. We may write

$\nu_1 = \frac{\Sigma f(x')}{N}$ = first moment of the distribution about the arbitrary origin.

$\nu_2 = \frac{\Sigma f(x')^2}{N}$ = second moment of the distribution about the arbitrary origin.

$\nu_3 = \frac{\Sigma f(x')^3}{N}$ = third moment of the distribution about the arbitrary origin.

$\nu_4 = \frac{\Sigma f(x')^4}{N}$ = fourth moment of the distribution about the arbitrary origin.

"Moment" is a familiar mechanical term for the measure of a force with respect to its tendency to produce rotation. The strength of this tendency depends, obviously, upon the distance of the point at which the force is exerted from the origin. The term is used in statistics in a quite analogous sense, the class frequencies being looked upon as the forces in question. The distance of each class mid-point from the origin is the factor of prime importance in this respect. The moments of a distribution about any origin may be computed by multiplying the frequency of each class by a given power of its distance, along the x-axis, from the origin, summing the resulting products and dividing by the number of cases. If the first moment is desired, the first power of the x-distance is employed; if the fourth moment, the fourth power of the x-distance, etc. The subscripts indicate the moments represented by the various symbols.

The most significant moments, for statistical purposes, are those which relate to the arithmetic mean as origin. Representing these moments by π, we have the following relationships:

First moment about the mean $= \pi_1 = 0$
Second " " " " $= \pi_2 = \nu_2 - \nu_1^2$
Third " " " " $= \pi_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3$
Fourth " " " " $= \pi_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4$

The computation of these moments from the data, as classified, involves the assumption that the items in each class can be treated as though they were concentrated at the mid-point of that class. It has been established that, under certain conditions, calculations made on this assumption are subject to a constant error. In particular, it has been shown that the values of the second and fourth moments are not the same, when computed from grouped data, as when computed from ungrouped data.

W. F. Sheppard[1] has worked out certain corrections for this bias. His corrections may be applied when two conditions prevail:

(1) When the distribution relates to a continuous variable.

(2) When the frequency curve is characterized by "high contact", i.e., when the frequency curve tapers off gradually in both directions.

The symbol μ is employed to represent a corrected moment about the mean. The application of Sheppard's corrections gives us the following final formulation:

$$\mu_1 = 0$$

$$\mu_2 = \pi_2 - \frac{1}{12}$$

$$\mu_3 = \pi_3$$

$$\mu_4 = \pi_4 - \frac{1}{2}\pi_2 + \frac{7}{240}$$

(In applying the corrections 1/12 and 7/240, the corresponding decimal values, .083333 and .029167, will generally be employed.) It is assumed in making these corrections that a class-interval unit has been employed in measuring deviations from the mean.

It may be noted in passing that the standard deviation is the square root of the second moment about the mean. For the uncorrected value,

$$\sigma = \sqrt{\pi_2}$$

[1] *Cf. Proceedings of the London Mathematical Society*, Vol. XXIX, 353–380.

If Sheppard's corrections are to be applied

$$\sigma = \sqrt{\mu_2}$$

The calculation of the moments of the frequency distribution of telephone subscribers is shown below. Sheppard's corrections are applied, since the curve is marked by reasonably high contact. It is a discontinuous distribution, but the unit (1) is so small in comparison with the range that it may be treated as continuous.

$$\nu_1 = \frac{-956}{995} = -.960804$$

$$\nu_2 = \frac{9676}{995} = 9.724623$$

$$\nu_3 = \frac{-22952}{995} = -23.067337$$

$$\nu_4 = \frac{283564}{995} = 284.988945$$

$$\pi_1 = 0$$

$$\pi_2 = \nu_2 - \nu_1^2 = 9.724623 - .923144 = 8.801479$$

$$\pi_3 = \nu_3 - 3\nu_1\nu_2 + 2\nu_1^3 = -23.067337 + 28.030370 - 1.773922 = 3.189111$$

$$\pi_4 = \nu_4 - 4\nu_1\nu_3 + 6\nu_1^2\nu_2 - 3\nu_1^4$$
$$= 284.988945 - 88.652760 + 53.863384 - 2.556586$$
$$= 247.642983$$

$$\mu_1 = 0$$

$$\mu_2 = \pi_2 - \frac{1}{12} = 8.801479 - .083333 = 8.718146$$

$$\mu_3 = \pi_3 = 3.189111$$

$$\mu_4 = \pi_4 - \frac{1}{2}\pi_2 + \frac{7}{240} = 247.642983 - 4.400739 + .029167$$
$$= 243.271411$$

Criteria of Curve Type

Having these values, we may return to a consideration of the main problem, the utilization of our knowledge of the normal curve. There are certain criteria which enable

us to determine readily whether a given distribution may be described by a curve of the normal type. These may be derived from the corrected moments of the given distribution.

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{10.170429}{662.632015} = .01534853$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{243.271411}{76.006070} = 3.200683$$

$$\kappa_2 = \frac{\beta_1(\beta_2 + 3)^2}{4(4\beta_2 - 3\beta_1)\ (2\beta_2 - 3\beta_1 - 6)}$$

$$= \frac{.01534853 \times 38.448470}{4(12.756686)\ (.355320)} = \frac{.5901275}{18.130823}$$

$$\kappa_2 = .032548$$

For the normal curve these criteria have the following values

$$\beta_1 = 0$$
$$\beta_2 = 3$$
$$\kappa_2 = 0$$

We may conclude, tentatively, that the normal curve may be used to describe the given distribution.[1]

Fitting a Normal Curve; Use of Table of Ordinates

The process of fitting the normal curve has been simplified by the preparation of tables which express ordinates of such a curve as fractional parts of the maximum ordinate, y_o. The following very short table will serve to indicate the nature of the procedure, though it is not detailed enough to employ in the actual computations.[2]

[1] It is possible that a better fit would be obtained with a Pearsonian Type IV curve, but, as will be seen, a reasonably good fit is secured with the normal curve, and this curve has an additional advantage in the matter of simplicity.

[2] A comprehensive table of ordinates of the normal curve, in terms of abscissa, will be found in Pearson's *Tables for Statisticians and Biometricians.*

TABLE 125

Ordinates of the Normal Curve

Expressed as Fractional Parts of the Maximum Ordinate

x/σ	y/y_o	x/σ	y/y_o
0.0	1.00000	2.5	.04394
0.1	.99501	2.6	.03405
0.2	.98020	2.7	.02612
0.3	.95600	2.8	.01984
0.4	.92312	2.9	.01492
0.5	.88250	3.0	.01111
0.6	.83527	3.1	.00819
0.7	.78270	3.2	.00598
0.8	.72615	3.3	.00432
0.9	.66689	3.4	.00309
1.0	.60653	3.5	.00219
1.1	.54607	3.6	.00153
1.2	.48675	3.7	.00106
1.3	.42956	3.8	.00073
1.4	.37531	3.9	.00050
1.5	.32465	4.0	.00034
1.6	.27804	4.1	.00022
1.7	.23575	4.2	.00015
1.8	.19790	4.3	.00010
1.9	.16448	4.4	.00006
2.0	.13534	4.5	.00004
2.1	.11025	4.6	.00003
2.2	.08892	4.7	.00002
2.3	.07100	4.8	.00001
2.4	.05614	4.9	.00001
		5.0	.00000

In the column x/σ deviations from the mean are expressed in units of the standard deviation. In the column y/y_o ordinates of the normal curve at given distances from the mean are given as fractional parts of the maximum ordinate. Thus the ordinate erected at a distance of one standard deviation from the mean is equal to $.60653y_o$ while an ordinate distant from the mean by 2.8σ is equal to $.01984y_o$.

In fitting a normal curve to a given distribution, therefore, it is necessary, first, to express the deviations from the mean in terms of σ; second, to determine the value of the maximum ordinate, and finally, using such a table as the one above, to compute all other necessary ordinates. The value of the maximum ordinate is secured from the relation

$$y_o = \frac{N}{\sigma\sqrt{2\pi}}$$

or, substituting the value of $\sqrt{2\pi}$, from

$$y_o = \frac{N}{2.506628\sigma}$$

In the present problem

$$N = 995$$

and

$$\begin{aligned}\sigma &= \sqrt{\mu_2} \\ &= \sqrt{8.718146} \\ &= 2.953\end{aligned}$$

Hence

$$\begin{aligned}y_o &= \frac{995}{2.506628 \times 2.953} \\ &= 134.42\end{aligned}$$

The value of the arithmetic mean is 476.96.

The subsequent calculations appear in the table on the following page.

In the application of the formula for y_o and in carrying through the calculations shown in Table 126, the standard deviation should be expressed in class-interval units. This is necessary in order that the computed frequencies may be comparable with the actual frequencies, as tabulated by classes.

TABLE 126

Computation of Ordinates of Normal Curve Fitted to Frequency Distribution of Telephone Subscribers

Number of messages m	*Deviation from M (in class intervals)* x	$\frac{x}{\sigma}$	$\frac{y}{y_o}$	*Computed frequency (ordinates)* y_c
25	−9.0392	−3.06	.00153	.21
75	−8.0392	−2.72	.02474	3.33
125	−7.0392	−2.38	.05888	7.91
175	−6.0392	−2.05	.12230	16.44
225	−5.0392	−1.71	.23176	31.15
275	−4.0392	−1.37	.39123	52.59
325	−3.0392	−1.03	.58834	79.08
375	−2.0392	−.69	.78817	105.95
425	−1.0392	−.35	.94055	126.43
475	−.0392	−.01	.99995	134.41
525	.9608	.32	.94856	127.51
575	1.9608	.66	.80429	108.11
625	2.9608	1.00	.60653	81.53
675	3.9608	1.34	.40747	54.77
725	4.9608	1.68	.24385	32.78
775	5.9608	2.02	.13000	17.47
825	6.9608	2.36	.06174	8.30
875	7.9608	2.70	.02612	3.51
925	8.9608	3.03	.01015	1.36
975	9.9608	3.37	.00342	.46
1025	10.9608	3.71	.00103	.14
1075	11.9608	4.05	.00027	.04
				993.48

TESTING THE GOODNESS OF FIT

The actual distribution and the normal curve, as fitted by the above computations, are plotted in Fig. 86. It is apparent, by inspection, that the normal curve gives a fairly good fit to the data, although there are several points at which the differences are marked. A natural question arises as to the reason for the failure of the normal curve to fit at all points. There are two possible answers

to such a question. The failure to fit may be due merely to chance fluctuations such as are found in any sample. We may have an underlying law of distribution of residence subscribers, classified by message use, which accords perfectly with the normal law of error, but the particular sample selected may be marked by certain irregularities

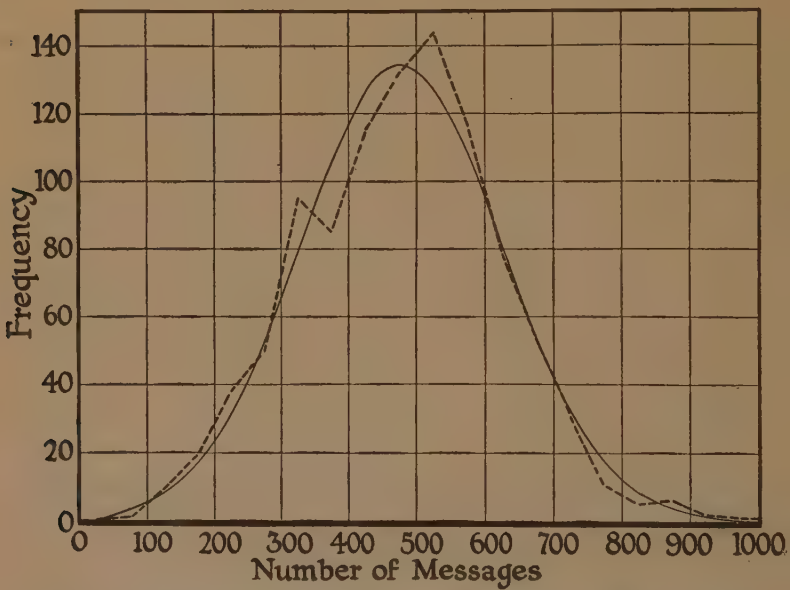


FIG. 86. — Illustrating the Fitting of a Normal Curve to Frequency Distribution of Telephone Subscribers, Classified According to Message Use

which would be ironed out if a very large number of cases were included. On the other hand, the differences may be due to a fundamental failure of such a distribution to accord with the normal law of error. Such a law may not describe the distribution of telephone calls, in which case the normal curve should not be employed.

Which of these explanations accounts for the differences noted? We may attempt to answer this question in two ways. We may consider the differences between actual and computed frequencies at several points where these

differences are greatest, and determine whether these differences are such as might be due to the fluctuations arising in random sampling; or we may seek to determine whether, for the curve as a whole, the combined discrepancies between observed and computed frequencies are such as might arise in random sampling, if the underlying distribution of telephone calls actually accords with the normal law of error. Both these methods may be employed.

It is necessary, however, to compute more exactly the theoretical frequencies. The ordinates of the normal curve do not measure these frequencies with sufficient accuracy. The ordinate at the middle of an interval one unit wide would accurately measure the area within that interval only if the curve which marked its upper boundary were straight. This is not true with respect to the segments under the normal curve. Accordingly, it is necessary to determine the theoretical frequencies by measuring areas instead of ordinates.

The Determination of Theoretical Frequencies

The entire area under a given frequency curve is taken to represent the total number of frequencies. Given information as to the proportion of the total area within a given segment, it would be easy to compute the frequencies represented by this segment. Fortunately such information is to be had, with respect to the normal curve, from prepared tables of the probability integral, of which the table on page 538 is an example. A much more detailed table is needed for accurate computation.[1]

Since the normal curve is symmetrical about the maximum ordinate, the values given below apply both to positive and negative values, as measured from the mean.

[1] Detailed tables of the probability integral as calculated by Dr. W. F. Sheppard have been widely circulated. Pearson's *Tables for Statisticians and Biometricians* contains such a table in excellent form for use.

TABLE 127

Area of the Normal Curve in Terms of Abscissa

(Giving fractional parts of the total area between y_o and ordinates erected at varying distances from y_o)

x/σ	a	x/σ	a
0.0	.00000	2.0	.47725
0.1	.03983	2.1	.48214
0.2	.07926	2.2	.48610
0.3	.11791	2.3	.48928
0.4	.15542	2.4	.49180
0.5	.19146	2.5	.49379
0.6	.22575	2.6	.49534
0.7	.25804	2.7	.49653
0.8	.28814	2.8	.49744
0.9	.31594	2.9	.49813
1.0	.34134	3.0	.49865
1.1	.36433	3.1	.49903
1.2	.38493	3.2	.49931
1.3	.40320	3.3	.49952
1.4	.41924	3.4	.49966
1.5	.43319	3.5	.49977
1.6	.44520	3.6	.49984
1.7	.45543	3.7	.49989
1.8	.46407	3.8	.49993
1.9	.47128	3.9	.49995
		4.0	.49997

In using such a table, deviations from the mean are first expressed in terms of the standard deviation. The proportion of the total area lying between any two ordinates may then be readily determined. For example: What proportion of the cases in a normal distribution lies between the maximum ordinate and an ordinate erected at a distance from the mean equal to $+\sigma$? Reading down the x/σ column to 1.0, we find the value .34134 opposite it. This, in ratio form, is the proportion of cases falling within the limits indicated. Expressing this ratio as a percentage, we have 34.134 per cent as the answer to our question.

Fig. 87 shows the relation of this area (the shaded area A) to the total area under the curve.

What proportion of the total number of cases in a normal frequency distribution will fall between an ordinate erected at a distance from the mean equal to -1.4σ and one erected at -2σ? From the table we find that 41.92 per cent of the total area will lie between y_0 and the ordinate at -1.4σ; 47.73 per cent will lie between y_0 and the

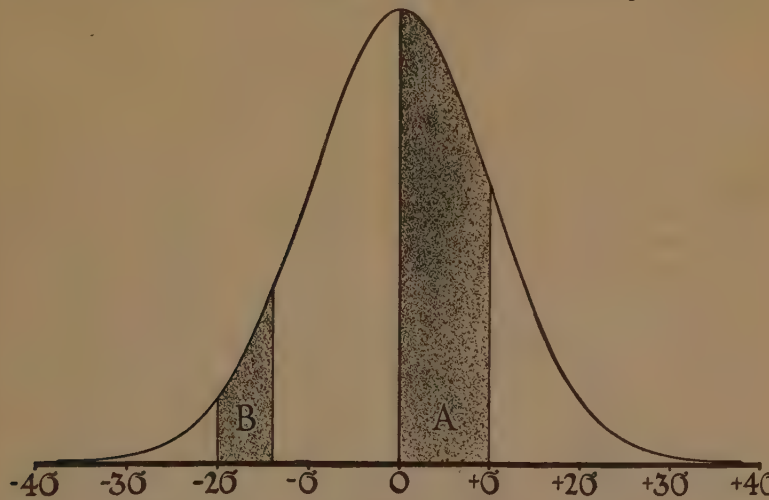


FIG. 87. — An Illustration of the Measurement of Areas Under the Normal Curve

ordinate at -2σ. The difference, 5.81 per cent, will fall between the ordinates at -1.4σ and at -2σ. This may be converted into actual frequencies by taking this proportion of the total number of cases in the given distribution. The shaded segment *B* in Fig. 87 represents the area thus marked off.

The procedure to employ in computing theoretical frequencies from such a table of areas will be clear from these examples. It is necessary only to find the area between the maximum ordinate and ordinates erected at the various class limits. By the simple process of subtraction the area within each class, and hence the theoretical frequencies, may be computed. The procedure is illustrated in the following table, relating to the distribution of telephone subscribers.

TABLE 128

Illustrating the Computation of Theoretical Frequencies from a Table of Areas

(1) *Class limit*	(2) *Deviation from mean* $\frac{x}{\sigma}$	(3) *Proportion of area between y_0 and ordinate at* $\frac{x}{\sigma}$	(4) *Number of cases between y_0 and ordinate at* $\frac{x}{\sigma}$	(5) *Theoretical frequencies, by classes*	
0	−3.23	.4993810	496.88		
50	−2.89	.4980738	495.58	0–50	1.92[1]
100	−2.55	.4946139	492.14	50–100	3.44
150	−2.22	.4867906	484.36	100–150	7.78
200	−1.88	.4699460	467.60	150–200	16.76
250	−1.54	.4382198	436.03	200–250	31.57
300	−1.20	.3849303	383.01	250–300	53.02
350	−.86	.3051055	303.58	300–350	78.63
400	−.52	.1984682	197.48	350–400	106.90
450	−.18	.0714237	71.07	400–450	126.41
500	+.16	.0635595	63.24	450–500	134.31
550	+.495	.1896931	188.74	500–550	125.50
600	+.83	.2967306	295.25	550–600	106.51
650	+1.17	.3789995	377.10	600–650	81.85
700	+1.51	.4344783	432.31	650–700	55.21
750	+1.85	.4678432	465.50	700–750	33.19
800	+2.19	.4857379	483.31	750–800	17.81
850	+2.53	.4942969	491.83	800–850	8.52
900	+2.87	.4979476	495.46	850–900	3.63
950	+3.20	.4993129	496.82	900–950	1.36
1000	+3.54	.4997999	497.30	950–1000	.48
1050	+3.88	.4999478	497.45	1000–1050	.15
1100	+4.22	.4999878	497.49	greater than 1050	.05
					995.00

[1] The theoretical distribution shows .62 of a case below -3.23σ. Since this would be meaningless in the present example, this amount is added to the theoretical frequency between 0 and 50.

THE STANDARD ERROR OF SAMPLING AS A TEST OF GOODNESS OF FIT

Now that we have this more accurate compilation of theoretical frequencies we are prepared to test the results

of the fitting process to ascertain whether the normal curve gives a reasonable fit. We may first take the greatest differences between the observed and the theoretical frequencies, and determine the significance of such discrepancies. As we have seen, the standard deviation may be computed, when the probabilities are known, from the formula

$$\sigma = \sqrt{npq}$$

where p refers to the probability of success and q to the probability of failure. In dealing with a frequency distribution we may replace p by $\frac{f}{N}$ (f representing the theoretical frequency at a selected point on the x-scale and N the total number of cases). In this case q will be equal to $\frac{N-f}{N}$. Inserting these measures in place of p and q in the general formula for the standard deviation, we have

$$\sigma_s = \sqrt{N \times \frac{f}{N} \times \left(\frac{N-f}{N}\right)}$$
$$= \sqrt{\frac{f(N-f)}{N}}$$

This measure is called the *standard error of sampling.*

The points on the scale, at which the differences between the observed and theoretical frequencies are the greatest, are the following:

(Theoretical frequencies are represented by f and observed frequencies by f_o.)

m	f_o	f	$f_o - f$
325	95	78.63	+16.37
375	85	106.90	−21.90
525	144	125.50	+18.50
775	11	17.81	−6.81

For the standard error of sampling in the first case we have

$$\sigma_s = \sqrt{\frac{78.63(995 - 78.63)}{995}} = 8.51$$

Computing similar measures for the three other classes, we secure the following values:

σ_s for class with midpoint 325 = 8.51
σ_s " " " " 375 = 9.73
σ_s " " " " 525 = 10.47
σ_s " " " " 775 = 4.18

We may now answer in a definite way the question as to whether the discrepancies between observed and theoretical frequencies would be possible if the data relating to telephone calls actually followed the normal law of error in its distribution. The difference in the case of the class with midpoint 325 is + 16.37; σ_s for this class is 8.51. The deviation is + 1.9 times σ_s. How often would such a deviation occur in actual practice if the underlying distribution were really normal?

We may answer this question by referring to the table of the probability integral. (Table 127.) Opposite the value $\frac{x}{\sigma} = 1.9$ we find .47128. This means that approximately 47 per cent of all the cases will lie between the maximum ordinate and an ordinate at 1.9σ from the mean. Approximately 94 per cent of the total area lies between ordinates erected at $\pm 1.9\sigma$ from the mean, hence the chances are about 6 out of 100 that a given value will differ from the mean by more than 1.9σ. This particular deviation, then, is one which might be accounted for by mere chance fluctuations of sampling, and does not prove that the normal curve is unsuited to the present distribution.

Similar tests may be applied to the differences found in the other classes. The greatest is that in the class with a mid-value of 375, in which the difference of − 21.90 represents $2.2\sigma_s$. Reference to the tables shows that a negative deviation of this amount or more would occur about 1.5 times out of 100. Counting both positive and negative deviations, such a difference between observed and theo-

retical frequencies might be expected about 3 times out of each 100 trials. We may conclude, then, that even the greatest discrepancy actually found in the present example is not so large as to prove that the normal curve is inapplicable in describing the distribution of residence subscribers on the basis of message use.

The Chi-Square Test of Goodness of Fit

It is obvious that a somewhat more conclusive result would be secured if we took account of all the classes and all the differences, instead of confining ourselves to the most exceptional cases. Karl Pearson has evolved such a measure, χ^2, which enables us to determine the goodness of fit of a given type of frequency curve, when all the data are taken into account. Having the theoretical and observed frequencies, this measure is computed quite readily, as is demonstrated in the following example:

Table 129

Computation of χ^2 for Testing Goodness of Fit
Normal Curve of Error Fitted to Distribution of Telephone Subscribers

(1) *Class limits*	(2) *Observed frequency* f_o	(3) *Theoretical frequency* f	(4) $(f_o - f)$	(5) $\frac{(f_o-f)^2}{f}$
150 and less	10	13.14	−3.14	.75
150–200	19	16.76	+2.24	.30
200–250	38	31.57	+6.43	1.31
250–300	50	53.02	−3.02	.17
300–350	95	78.63	+16.37	3.41
350–400	85	106.90	−21.90	4.49
400–450	115	126.41	−11.41	1.03
450–500	132	134.31	−2.31	.04
500–550	144	125.50	+18.50	2.73
550–600	116	106.51	+9.49	.85
600–650	79	81.85	−2.85	.10
650–700	54	55.21	−1.21	.03
700–750	31	33.19	−2.19	.14
750–800	11	17.81	−6.81	2.60
More than 800	16	14.19	+1.81	.23
	995	995.00	15 Groups	$\chi^2 = 18.18$

It will be noted that in the construction of this table the three classes at the lower end of the distribution have been lumped into one, and that the same thing has been done with the six classes at the upper end of the distribution. This is done to avoid unduly magnifying slight differences between the tails of the observed and theoretical distributions.

The value of χ^2 furnishes us with the required measure of goodness of fit, though its interpretation necessitates the use of detailed tables.[1] The following extract from such a table will indicate the procedure:

Values of P for given values of χ^2 and n'

χ^2	$n' = 14$	$n' = 15$	$n' = 16$
16	.249129	.313374	.382051
17	.199304	.256178	.318864
18	.157520	.206781	.262666
19	.123104	.164949	.213734
20	.095210	.130141	.171932

In this table χ^2 represents $\Sigma\left(\frac{(f_o - f)^2}{f}\right)$

and n' represents the number of classes. In the present case

$$\chi^2 = 18.18$$
$$n' = 15$$

For $n' = 15$ and $\chi^2 = 18$, the value of P is .206781; for $n' = 15$ and $\chi^2 = 19$, $P = .164949$. Interpolating for the given value of χ^2 we find that $P = .199$.

P measures the probability that, if the real distribution were in accord with the normal law of error, we should obtain a fit as bad or worse than that actually secured. That is, in 199 out of 1000 trials, or about 1 out of 5, we should secure a fit as bad or worse than the one we have, assuming the underlying distribution to be normal. These are not at all excessive odds, and we may conclude that the

[1] These tables, as worked out by W. Palin Elderton, are to be found in Karl Pearson's *Tables for Statisticians and Biometricians*, 26–30.

normal curve gives such a reasonably good fit that it may be employed as a means of describing the given distribution.

Having reached this conclusion, all the information which is had with respect to distributions following the normal law of error is immediately available in the study of the given distribution. In using the original frequency table, we are limited to the classes there established. We may now go beyond this and determine how many cases may be expected between any two values on the x-scale. We may compute the probability of a case falling anywhere on the scale, or above or below any given value. And not only is our knowledge of details greater. In so far as we are assured of the representative character of our sample we have a basis for generalization which would be afforded by no amount of study of the particular distribution as a thing apart.[1]

Note on the Description of the Frequency Distribution

With the aid of the criteria explained in this chapter it is possible to describe a frequency distribution somewhat more accurately, in some respects, than could be done with the measures employed in the earlier chapters. A treatment of this subject is beyond the scope of the present book, but it seems advisable to indicate briefly the nature of these additional measures.

The value of β_2 serves as a measure of the degree of "flat-toppedness" found in a given curve. If $\beta_2 = 3$, as in the normal type, the curve is said to be *mesokurtic*. If $\beta_2 < 3$ the curve is *platykurtic*, or flatter than the normal type. If $\beta_2 > 3$, as in the example given above, the curve is *leptokurtic*, or more peaked than the normal.

[1] As was stated, the normal curve is but one type of frequency curve, though one of basic importance. A comprehensive system of frequency curves is that associated with the name of Karl Pearson, who has derived equations to and has described in detail a number of standard types. An account of other fundamental types will be found in the books by Arne Fisher referred to at the end of this chapter. A treatment of this subject in detail is beyond the scope of the present work.

A measure of skewness which is more accurate than those given early in the book may also be computed from these measures. Karl Pearson has shown that the quantity

$$\chi = \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$$

serves as a measure of the degree of asymmetry of a given curve.[1] Inserting the values of β_1 and β_2 given above we have, in the case of the distribution based on message use,

$$\chi = -.05558$$

(χ is positive if the mean is greater than the median, negative if the mean is less than the median. In the present case the value of the mean is 476.96, that of the median is 482.39, hence the skewness is negative.)

Finally, the distance, d, between the mean and the mode may be determined from the relation

$$d = \chi \times \sigma$$

In the distribution described above (relating to telephone use) σ, in original units, equals 147.65. Hence

$$d = -.05558 \times 147.65 = -8.21$$

Since

$$M_o = M - d$$

we have

$$M_o = 476.96 + 8.21 = 485.17$$

This gives a truer approximation to the modal value than any of the methods discussed in Chapter IV.

REFERENCES

Bowley, Arthur L. *Elements of Statistics* (259–286).
Brunt, David. *The Combination of Observations* (11–28).
Carver, H. C. *Frequency Curves* (In Rietz, H. L., *Handbook of Mathematical Statistics*, 92–119).
Elderton, W. P. *Frequency Curves and Correlation.*
Fisher, Arne. *An Elementary Treatise on Frequency Curves. The Mathematical Theory of Probabilities.*

[1] χ, the measure of skewness, is quite distinct from χ^2, which is computed in testing goodness of fit. The two measures should not be confused.

KELLEY, TRUMAN L. *Statistical Method* (94–108). (The Kelley-Wood Table of the Normal Probability Integral is given as an appendix.)

PEARL, RAYMOND. *Medical Biometry and Statistics* (220–263).

PEARSON, KARL. *Tables for Statisticians and Biometricians.* (The *Introduction* to these tables will be found particularly useful.)

SHEPPARD, W. F. *On the Calculation of the Most Probable Values of Frequency Constants for Data Arranged According to Equi-distant Divisions of a Scale.* Proceedings of the London Mathematical Society, Vol. XXIX, 1898.

The Calculation of the Moments of a Frequency Distribution. Biometrika, Vol. V.

YULE, G. U. *An Introduction to the Theory of Statistics* (291–316).

CHAPTER XVI

STATISTICAL INDUCTION AND THE PROBLEM OF SAMPLING

The preceding pages have been devoted to an account of the tools which are employed in statistical analysis. Examples illustrating the application of these tools to certain specific problems have been presented, but the emphasis throughout has been on technique. It is appropriate at this point that we stand off a distance, enlarging our perspective, and consider certain general problems relating to the application of these tools. What is their proper place in economic and business research? What are the assumptions involved in using them and what are their limitations? These are questions which cannot be ignored, though detailed treatment is not possible within the limits of the present volume.

Statistical Description and Statistical Induction

In approaching this subject we must first make clear the distinction between *statistical description* and *statistical induction*. By employing the methods of statistics it is possible, as we have seen, to describe succinctly a mass of quantitative data. Hundreds or thousands of individual cases may be classified, and a frequency distribution formed. The essence of this distribution may be boiled down to perhaps four measures — of central tendency, of variation, of skewness and of kurtosis. A tremendous gain has been realized in thus replacing the multiplicity of individual cases by a limited number of measures which

describe the characteristics of the group as a whole. The possession of such tools makes it possible for our limited powers of perception to grasp the significance of facts in the mass. Again, the methods of statistics enable us to describe relations between variable quantities. By securing the equation to an appropriate curve fitted to the data by mathematical methods, we are enabled to determine how much, on the average, one quantity changes in value as one or more related quantities vary. This may be supplemented by a measure of the scatter or dispersion about the fitted curve, and by a final measure, in abstract terms, of the degree of correlation between the dependent and the one or more independent variables.

In so far as the results are confined to the cases actually studied, these various statistical measures are merely devices for describing certain features of a distribution, or certain relationships. Within these limits the measures may be used with perfect confidence, as accurate descriptions of the given characteristics. But when we seek to extend these results, to generalize the conclusions, to apply them to cases not included in the original study, a quite new set of problems is faced.

The logical process by which general conclusions are drawn from a study of particular cases is termed *induction*, as opposed to *deduction*, which involves the drawing of conclusions from certain general principles. By *statistical induction* or *statistical inference* is meant the generalization of statistical results, on the assumption that a given statistical measure may be taken to apply to a larger group than that from which it was actually derived. We are employing this procedure constantly in practical statistical work, though not always with a full realization of the assumptions inherent in that process and of the limitations which attach to it. These assumptions and limitations may be briefly considered.

The Generalization of Statistical Results

The problem at issue in considering the validity of statistical induction may be put in the following form: A statistical measure — an average, a frequency ratio, a coefficient of correlation — has been derived from the study of certain data drawn from a large population. (The term "population" refers to a group of things or phenomena having, presumably, certain characteristics in common.) May we assume that, if additional samples were taken from the same population, the corresponding measures would have the same values? If not, may we determine, from the results secured in the given case, the approximate limits to the fluctuations to be expected in these measures as derived from successive samples? Here, obviously, is a problem of supreme importance. Karl Pearson has called it "the fundamental problem of practical statistics." If we cannot be assured of a certain degree of stability in the results secured from successive samples it would be quite invalid to generalize from the study of a limited number of cases. No weight would attach to any study except one which covered the entire universe of things or phenomena composing the given population. Yet such all-inclusive studies of economic phenomena are practically impossible. Index numbers of prices, of wages, of living costs, equations describing the relation between the production and price of given commodities, coefficients of correlation between temperature and crop yield — all must of necessity be based on the study of samples.

The Assumption of Uniformity in Nature

The statistical measures secured from successive samples might be assumed to be stable if the validity of two prior assumptions be granted. The first of these is, in general terms, an assumption of uniformity in nature, the assumption that there is, in nature, a "limitation to the amount

of independent variety." When dealing with quantitative data this uniformity in nature appears in the stability of large numbers, as exemplified by the curious regularity in such phenomena as birth rates or death rates. Nature, in other words, is not marked by utter chaos; principles of regularity, order and stability appear in all natural processes, and these principles are strongly evident when we deal with masses of quantitative data. Therefore, when we generalize such a measure as an index number of wholesale prices, we do so on some such assumption as this: It is reasonable to suppose that, in the larger population to which this result is to be applied, there exists a uniformity with respect to the characteristic or relation we have measured. As a result of this uniformity, we should expect statistical measures derived from successive samples drawn from this population to fluctuate within definite and assignable limits, which we may approximate in advance.

It is evident that in making this assumption, in saying "It is reasonable to suppose . . . ," we are introducing an hypothesis which is incapable of complete verification by purely statistical methods. There is, thus, in every statistical induction, an *a priori* element. The statistical conclusion can never stand completely on its own feet. It must be endorsed by reason and judgment if it is to carry conviction. If a high positive coefficient of correlation were secured from the study of a sample relating to banana importations and the suicide rate, this would not furnish convincing evidence of a causal relation, or a relation of contingency, between these two variables. There would be no reasonable basis for assuming that, in the larger universe of phenomena from which the sample was drawn, there would be uniformity with respect to this relationship.

The Necessity of a Representative Sample

The second condition upon which the validity of this type of reasoning depends is implicit in the preceding argument. This is the assumption that the sample from which our first results were derived is thoroughly representative of the entire population to which the results are to be applied. The securing of such a representative sample is a first condition of valid statistical induction.

How may a representative sample be secured? It is generally laid down as essential that the members included in the sample shall be *random* members of the population at large, that in the selection of the sample there shall be present no element of preference or bias which would tend toward the inclusion or exclusion of certain members of the larger group. Each member of the population should have an equal chance of inclusion in the sample on which the induction is to be based. This general requirement, as J. M. Keynes has pointed out, should be interpreted to mean that *with respect to the generalization in question* the members of the sample should be random members of the population at large. Great care is generally needed in securing a purely random selection. The obvious procedure of picking the most readily available cases would by no means meet the condition of random selection. Certain important elements in the universe of facts to which the conclusions are to be applied may be excluded through the play of an unconscious bias unless careful attention is given to the selection of cases.

Conditions of Simple Sampling

G. U. Yule[1] has named certain other conditions which are tacitly assumed in deducing formulas relating to the sampling process. The assumptions upon which the process of "simple sampling" rests are:

[1] *Introduction to the Theory of Statistics*, 5th ed., 259–261.

1. An assumption of complete independence between the several events composing the sample. Thus to the extent that a change in the price of one commodity is affected by a change in the price of another, the two price changes are not independent, and their inclusion in a single sample violates the conditions of simple sampling.

2. An assumption that there is no essential difference between the localities from which the members of the sample are taken, and that there has been no essential change in underlying conditions during the period of time covered by the observations.

3. An assumption that "the conditions that regulate the appearance of the character observed (are) not only the same for every sample, but for every individual in every sample."

When all these conditions of sampling have been observed, it is possible to assign in advance limits within which we may expect statistical measures derived from different samples of the same population to fluctuate. This means that we may apply to the population at large statistical measures secured from the study of a sample not with confidence in their perfect stability, but with fairly definite knowledge of the margin of error involved in thus extending our results. Where the necessary conditions are fulfilled statistical induction is a valid procedure.

The Significance of Measures of Reliability

The nature of the measures of stability which we seek should be clearly understood. When we generalize from a knowledge of certain statistical quantities we do so by applying to a larger group the value for mean, standard deviation, or coefficient of correlation which has been computed from a sample. And we wish to determine the limits within which these results would fluctuate if computed from a number of different samples drawn from the same population. A measure of these limits will serve as a measure of the reliability of the given results, when extended.

Such a measure might be secured by the laborious process of studying a great many different samples, just as the dice were thrown 4096 times. Thus we might desire to test the reliability of an average of weekly earnings of a certain class of workers. A first average might be secured from a sample composed of 250 individual records. This result might be tested by computing 499 additional samples, each based on 250 individual records. These 500 averages would not be identical in value, but if they were tabulated a frequency distribution closely approximating the normal type would be secured. From this distribution we might compute the mean of all the averages and the standard deviation of these averages. This standard deviation would serve as a measure of the variation found in the average of weekly earnings, as computed from successive samples.

But it is generally impracticable to take 400 or 500 successive samples in order to test the reliability of a given measure. When the conditions of simple sampling, as set forth above, are fulfilled, it is possible to compute such measures of reliability more directly.

In the preceding chapter the results of an experiment involving the throwing of 12 dice 4096 times were given. A throw of 4, 5 or 6 was counted a success. The standard deviation, as computed from the results, was 1.712. But we found it possible, knowing the probabilities involved and the number of dice thrown each time, to compute a theoretical standard deviation from the formula

$$\sigma = \sqrt{npq}$$

This had a value of 1.732.

By an analogous procedure it is possible to compute similar quantities which serve as measures of the stability of statistical results, in so far as the fluctuations due to sampling are concerned. The formulas for these various measures of reliability, or standard errors, may be given

without attempting to indicate the processes by which they have been derived. The references at the end of this chapter may be consulted for the details of these processes.

Standard Errors of the Chief Statistical Measures

The reliability of an arithmetic mean, computed from a given sample, depends upon the value of the standard deviation of the original data and upon the number of cases included in the sample. The definite relationship is

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$

(σ, without subscript, is taken to refer to the standard deviation of the sample). This statement expresses the reliability of a given result in terms of the *standard error*. The *probable error* is more commonly used. The probable error of a normal distribution, it will be recalled, is always $.6745\sigma$. Thus we have

$$P.E._{Mean} = .6745 \frac{\sigma}{\sqrt{N}}$$

The meaning of the expression may be made clear by the use of an example. Assume that the mean weekly wage of 100,000 industrial workers in a given city is desired. A study including all workers is impossible, and a representative sample of 900 is selected. The mean weekly wage is found to be $27.50. May this be accepted as accurate? By how much would the results secured from additional samples be expected to differ from this value? Assuming that the standard deviation of the data is found to be $2.00, we have

$$P.E._{Mean} = .6745 \frac{\$2.00}{\sqrt{900}}$$
$$= \$.045$$

In reporting the result of the original study the mean weekly wage would be given as $27.50 \pm .045$. This is

interpreted to mean that if an additional sample of the same size were taken the chances would be even that the result would not differ from \$27.50 by more than \$.045.

In the same way measures of the accuracy of other statistical measures may be secured. The following are the formulas used in deriving the most important of these. (In each case the probable error may be derived by multiplying the standard error by .6745.)

$$\sigma_{Md} = 1.25331 \frac{\sigma}{\sqrt{N}} \qquad P.E._{Md} = .84535 \frac{\sigma}{\sqrt{N}}$$

$$\sigma_{Q1} = 1.36263 \frac{\sigma}{\sqrt{N}} \qquad P.E._{Q1} = .91908 \frac{\sigma}{\sqrt{N}}$$

(These measures have the same values for Q_3 as for Q_1.)

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2N}}$$

This formula for the standard error of the standard deviation holds exactly only when the distribution of observations in the sample is normal. For a distribution of any type the reliability of σ may be determined from

$$\sigma_\sigma = \sqrt{\frac{\mu_4 - \mu_2^2}{4\mu_2 \cdot N}}$$

The reliability of the coefficient of correlation, for a normal distribution, is measured by

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}}$$

This formula holds also for the coefficient of multiple correlation, coefficients of partial correlation and, approximately, for the correlation ratio.

The standard error of a coefficient of regression, assuming the distribution to be normal, is determined from the formula

$$\sigma_{b_{12}} = \frac{\sigma_1 \sqrt{1 - r^2}}{\sigma_2 \sqrt{N}}$$

This form may be adapted to a coefficient of regression of any order.

As a test for linearity we have been given

$$\zeta = \eta^2 - r^2$$

But we wish to know whether, in a given case, the difference between η^2 and r^2 is due merely to a chance fluctuation of sampling, or is due to a real departure of the underlying relationship from the linear form. As the standard error of ζ Blakeman has proposed

$$\sigma_\zeta = 2\sqrt{\frac{\zeta}{N}}\sqrt{(1-\eta^2)^2 - (1-r^2)^2 + 1}$$

The use of this measure may be illustrated with reference to the problem relating to wheat yield which was considered in an earlier chapter. For the relation between wheat yield and amount of nitrogen used as fertilizer, we had

$$r = +\ .793$$
$$\eta = .964$$

Therefore

$$\zeta = \eta^2 - r^2 = .300$$

Inserting the given values in the formula

$$\sigma_\zeta = 2\sqrt{\frac{\zeta}{N}}\sqrt{(1-\eta^2)^2 - (1-r^2)^2 + 1}$$

and solving, we have

$$\sigma_\zeta = .074$$

With ζ having a value of .300, about 4.05 times its standard error, .074, there can be no question as to the non-linearity of the relationship. The difference between η^2 and r^2 is one which could hardly be due to chance fluctuations of sampling.

For the normal curve the standard error of χ, the measure of skewness, is derived from the formula

$$\sigma_\chi = \sqrt{\frac{3}{2N}}$$

For the distribution relating to telephone use given in the preceding chapter, we have

$$\sigma_\chi = \sqrt{\frac{3}{1990}} = .03883$$

The value of χ was found to be $-.05558$, about 1.43 times its standard error. This value of χ, therefore, might well be due to fluctuations of sampling, and does not prove that the true distribution is other than normal.

The standard error of the quantity d, which measures the distance between the mean and the mode, may be determined, for the normal curve, from the formula

$$\sigma_d = \sqrt{\frac{3}{2N}}\,\sigma$$

For the distribution of telephone calls this becomes

$$\sigma_d = \sqrt{\frac{3}{1990}}\,147.65$$

$$= 5.733$$

Since d had a value of 8.21, less than one and one-half its standard error, we are led to conclude that the difference between the mean and the mode is one which might easily be due to chance fluctuations of sampling. If d were more than three times its standard error it would indicate a difference between mean and mode which was probably not due to chance alone, but represented a fundamental departure of the distribution from the normal type.

A practical problem of some importance is frequently encountered in comparing two averages. If two samples are taken and the averages computed, a question arises as to whether the difference between these samples might arise from mere fluctuations in sampling, or whether it indicates a real difference between the two larger groups from which the samples are drawn. The formula employed in determining the standard error of the difference between two means is

$$\sigma_D = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$$

where σ_1 and σ_2 represent the standard deviations of the two samples, while N_1 and N_2 represent the numbers in the two samples. In using this measure, the actual difference between two averages is compared with the standard error of this difference. If the actual difference is greater than three times the standard error it is highly unlikely that the difference is due to fluctuations of sampling. Such a condition would argue a fundamental difference between the two groups from which the samples were drawn.

The formula for the standard deviation of sampling, which may be used in comparing theoretical and actual frequencies, is

$$\sigma_s = \sqrt{\frac{f(N-f)}{N}}$$

Its use has been explained in the preceding chapter.

Most of the measures of reliability presented above have been given in terms of the standard error rather than of the probable error. The latter may be derived readily from the former, as explained, but the standard errors seem more appropriate for general use. The most elaborate tables of the probability integral are those in which abscissal distances are measured in terms of the standard deviation. The significance of any given standard error, in terms of probabilities, may be readily determined from such tables.

Limitations to Measures of Reliability

Such measures as those for which the formulas were given above should be used with a full understanding of their limitations. In the first place, their significance depends upon the presence of a sufficiently large number of cases in the sample. If N falls below 15 in any case the above formulas for standard errors should not be applied, while a much larger number of cases is advisable if much

weight is to be attached to the results. For the coefficient of correlation 25 cases may be considered a minimum. The assumption is made in the interpretation of these standard errors that statistical measures secured from successive samples would be distributed in accordance with the normal law of error. When the number of cases is large this is approximately true, even though the original data are not so distributed. But with a small number of cases this assumption may be quite invalid, and no precise significance, in terms of probabilities, may be attached to the standard error.

Moreover, and this is a most important warning, the standard errors can be assumed to measure only errors arising from the fluctuations of simple sampling. This assumption injects certain elements of doubt into any statistical induction. We cannot, in the first place, be sure that the conditions of simple sampling, as given above, are actually fulfilled in a given case. They are rarely, if ever, perfectly fulfilled in the handling of economic data. Secondly, the standard errors as derived above can give no indication of the possibility of fluctuations in successive samples due to causes other than those arising from simple sampling. Fluctuations due to bias, due to the absence of random selection in the sampling process, due to persistent errors of any sort, quite elude this method of determining probable stability. These precautions must be constantly borne in mind in using these measures if the results of statistical study are to be properly interpreted and properly applied.

So serious are these limitations to the employment of the usual measures of probable error in connection with economic data that it would seem generally advisable to subordinate such measures to actual statistical tests of stability. By the study of successive samples, and by the testing of the subordinate elements in a given sample when broken up into significant sub-groups, much more may be

learned as to the reliability of a given measure and as to the possibility of applying it generally than by unquestioning acceptance and uncritical employment of the usual mathematical formulas for probable errors.

REFERENCES

Bowley, Arthur L. *Elements of Statistics* (312–342).

Broad, C. D. *On the Relation between Induction and Probability.* Mind. N. S. Vol. 27, 1918, and Vol. 29, 1920.

Editorial. *On the Probable Errors of Frequency Constants.* Biometrika. Vol. 2 (273–281).

Elderton, W. P. *Frequency Curves and Correlation* (131–138).

Fisher, Arne. *The Mathematical Theory of Probabilities.*

Kelley, Truman L. *Statistical Method* (94–108).

Keynes, J. M. *A Treatise on Probability.*

Mills, Frederick C. *On Measurement in Economics* (in Tugwell, R. G. ed. *The Trend of Economics,* 37–70).

Pearl, Raymond. *Medical Biometry and Statistics* (209–219).

Pearson, Karl. *The Fundamental Problem of Practical Statistics.* Biometrika. Vol. 13.

Rietz, H. L. *Random Sampling* (in Rietz, H. L. ed. *Handbook of Mathematical Statistics,* 71–81).

Whitaker, E. T. and Robinson, G. *The Calculus of Observations* (164–208).

Yule, G. U. *An Introduction to the Theory of Statistics* (254–290, 335–356).

APPENDIX A

THE METHOD OF LEAST SQUARES AS APPLIED TO CERTAIN STATISTICAL PROBLEMS

The method of least squares in the case of a single unknown quantity is merely a procedure for obtaining the most probable value of that quantity from a number of separate observations. The most probable value is that for which the sum of the squares of the deviations (or residuals) is a minimum. This is the arithmetic mean of the observations.

Where the measurements or observations do not relate directly to a single unknown quantity, but to *functions* of a number of unknown quantities, the problem is somewhat different. In the first case mentioned each observation is in the form of a single magnitude. In the present case each observation is in the form of an *observation equation* in which the observed values of the variables, as found in combination, are entered. The unknown quantities are the constants which define the functional relationship between the variables in question. Our problem is that of finding the most probable values of these constants, the true values being unknown.

As in the simpler case the most probable values are those for which the sum of the squares of the residuals is a minimum. In this case, however, the residuals are deviations, not from a single magnitude, as in the case of the arithmetic mean, but from the curve which describes the most probable functional relationship. In other words, the residuals are the differences between the computed and the actual values of the dependent variable.

Derivation of the Normal Equations

Representing by Y an observed value of the dependent variable, by Y_c the corresponding computed value, by v the residual, or difference between Y and Y_c, and by W_1, W_2, W_3 and W_4 different independent variables (or different functions of a single independent variable), we may write

$$\begin{aligned} Y_c &= f(W_1, W_2, W_3, W_4) \\ v &= Y_c - Y \\ &= f(W_1, W_2, W_3, W_4) - Y \\ \Sigma(v^2) &= \Sigma[f(W_1, W_2, W_3, W_4) - Y]^2 \end{aligned}$$

If the function in a particular case is of the type

$$Y_c = aW_1 + bW_2 + cW_3 + dW_4$$

we have

$$\Sigma(v^2) = \Sigma[(aW_1 + bW_2 + cW_3 + dW_4) - Y]^2$$

Our problem is that of determining the most probable values of the constants which define the function. These constants are represented, in the present case, by a, b, c and d. (The W's, it should be noted, refer to quantities which are known, once the observation equations are given. In the usual case the W's are different functions of a single variable, but this is not essential.) On the assumption that the errors of observation are distributed in accordance with the normal law of error, it may be demonstrated that the most probable values of a, b, c and d, in the above equation, are those which render $\Sigma(v^2)$ a minimum; i.e.,

$$\Sigma[(aW_1 + bW_2 + cW_3 + dW_4) - Y]^2 = \text{a minimum} \qquad (A)$$

The normal equations necessary for the solution may be obtained by equating to zero the partial derivatives of the above expression with respect to the unknowns, a, b, c and d. That is, we first differentiate the above function with respect to a, holding b, c and d constant, then with respect to b, holding a, c and d constant, then with respect

to d, holding a, b and c constant. Carrying through this operation with respect to a, we have

$$\frac{\partial}{\partial a}\Sigma[(aW_1 + bW_2 + cW_3 + dW_4) - Y]^2 = 0$$

or

$$\text{I} \quad \Sigma W_1[(aW_1 + bW_2 + cW_3 + dW_4) - Y] = 0$$

Differentiating equation (A) now with respect to b, we have

$$\frac{\partial}{\partial b}\Sigma[(aW_1 + bW_2 + cW_3 + dW_4) - Y]^2 = 0$$

or

$$\text{II} \quad \Sigma W_2[(aW_1 + bW_2 + cW_3 + dW_4) - Y] = 0$$

Differentiating equation (A) with respect to c,

$$\frac{\partial}{\partial c}\Sigma[(aW_1 + bW_2 + cW_3 + dW_4) - Y]^2 = 0$$

or

$$\text{III} \quad \Sigma W_3[(aW_1 + bW_2 + cW_3 + dW_4) - Y] = 0$$

Differentiating equation (A) with respect to d,

$$\frac{\partial}{\partial d}\Sigma[(aW_1 + bW_2 + cW_3 + dW_4) - Y]^2 = 0$$

or

$$\text{IV} \quad \Sigma W_4[(aW_1 + bW_2 + cW_3 + dW_4) - Y] = 0$$

The most probable values of the quantities a, b, c and d are secured by solving simultaneously the four normal equations thus obtained (numbered above I, II, III, IV).

Formation of the Normal Equations

When the observation equations are all of the first degree (i.e., of the first degree with respect to the unknown quantities, a, b, c, etc.) the normal equations may be secured by the following process:

1. Write the equation which describes the assumed relationship. The observation equations are derived by substituting in this equation the observed values of the variables, as found in combination.

2. Multiply each observation equation by the coefficient of the first unknown in that equation; the sum of the resulting equations constitutes the first normal equation.

3. Multiply each observation equation by the coefficient of the second unknown in that equation; the sum of the resulting equations constitutes the second normal equation.

Continue this process until normal equations equal in number to the unknown quantities are obtained.

The actual process of forming the normal equations in curve fitting may be simplified, and the writing out of the separate observation equations avoided, as was demonstrated in earlier sections. The following may be laid down as general rules for the formation of the desired normal equations:

1. Write the equation of the curve to be fitted. For the purpose of this explanation we may employ the general form

$$Y = aW_1 + bW_2 + cW_3 + dW_4 + \ldots \qquad (1)$$

where Y represents the dependent variable, a, b, c, d . . . represent the constants in the equation (the unknown quantities in the present instance) and W_1, W_2, W_3, W_4 . . . represent the coefficients of these unknowns. It is assumed that these coefficients represent variables, and that term is used with reference to them. Call this equation (1).

2. Multiply each term in equation (1) by the coefficient of the first unknown in (1) (i.e., by W_1) and place the summation sign, Σ, before each variable. This is the first normal equation (I).

3. Multiply each term in equation (1) by the coefficient of the second unknown (i.e., by W_2), and place the summation sign before each variable. This is the second normal equation (II).

4. Multiply each term in equation (1) by the coefficient of the third unknown (i.e., by W_3) and place the summation sign before each variable. This is the third normal equation (III).

5. Multiply each term in equation (1) by the coefficient of the fourth unknown (i.e., by W_4) and place the summation sign before each variable. This is the fourth normal equation (IV).

The process may be continued until normal equations equal in number to the unknown quantities are obtained.[1]

A Standard Set of Normal Equations

As a set of generalized normal equations secured by the above process and applying to any equation which can be put in the form

$$Y = aW_1 + bW_2 + cW_3 + dW_4 + \cdots$$

we have

I $\Sigma(W_1Y)$
$= a\Sigma(W_1^2) + b\Sigma(W_1W_2) + c\Sigma(W_1W_3) + d\Sigma(W_1W_4) + \cdots$

II $\Sigma(W_2Y)$
$= a\Sigma(W_1W_2) + b\Sigma(W_2^2) + c\Sigma(W_2W_3) + d\Sigma(W_2W_4) + \cdots$

III $\Sigma(W_3Y)$
$= \Sigma(W_1W_3) + b\Sigma(W_2W_3) + c\Sigma(W_3^2) + d\Sigma(W_3W_4) + \cdots$

IV $\Sigma(W_4Y)$
$= a\Sigma(W_1W_4) + b\Sigma(W_2W_4) + c\Sigma(W_3W_4) + d\Sigma(W_4^2) + \cdots$

By substituting for W_1, W_2, W_3, W_4, etc., the particular functions employed in a given case, these equations may be readily adapted to any type of curve in the fitting of which the method of least squares is applicable. Thus in fitting a curve represented by the equation

$$Y = a + bX + cX^2 + dX^3$$

substitutions in the standard normal equations given above are based upon the following relations:

$$W_1 = 1$$
$$W_2 = X$$
$$W_3 = X^2$$
$$W_4 = X^3$$

The changes to be made in the normal equations are obvious. $\Sigma(W_1Y)$ becomes $\Sigma(Y)$; $\Sigma(W_1^2)$ is equivalent

[1] These rules represent an adaptation of a similar series formulated by Raymond Pearl in *Medical Biometry and Statistics*, 341.

to $\Sigma(1^2)$, which is equal to N, the total number of observations. The first normal equation becomes

$$\Sigma(Y) = Na + b\Sigma(X) + c\Sigma(X^2) + d\Sigma(X^3)$$

The other normal equations are modified correspondingly.

In the example just given, the coefficients are all different functions of a single independent variable, X. It is not, of course, essential to the method of least squares that this be so. The coefficients, W_1, W_2, W_3, etc., may represent a number of independent variables, as in the case of multiple correlation.

The limitations to the method of least squares must be borne in mind in making use of it. This method, in its direct application, is limited to cases in which the equation to the curve to be fitted is linear in the constants, i.e., the observation equations must all be linear as regards the unknown values, a, b, c, etc. (This does not mean, of course, that the equation to the fitted curve must be linear.) As an example of this limitation, we may cite a curve having as equation $y = ab^{c^x}$, which cannot be fitted directly by the method of least squares. If the observation equations are non-linear they may be reduced to the linear form in many instances by the use of logarithms, and the method of least squares then employed.

Derivation of the Formula for the Standard Error of Estimate

It has been pointed out in the body of the text that the standard error of estimate may be derived as a by-product of the method of least squares. A more complete demonstration of this process may be given at this point.

When the partial derivative of the expression

$$\Sigma[(aW_1 + bW_2 + cW_3 + dW_4) - Y]^2 = \text{a minimum}$$

is equated to zero, with respect to the first unknown, a, we have

$$\Sigma W_1[(aW_1 + bW_2 + cW_3 + dW_4) - Y] = 0$$

Since

$$aW_1 + bW_2 + cW_3 + dW_4 - Y = v$$

we have as a necessary condition of fitting

$$\Sigma(vW_1) = 0$$

When the partial derivative of the same expression with respect to b is equated to zero, we have

$$\Sigma W_2[(aW_1 + bW_2 + cW_3 + dW_4) - Y] = 0$$

or, making the same substitution as in the preceding case,

$$\Sigma(vW_2) = 0$$

Repeating the operation with respect to c and d, we may show that

$$\Sigma(vW_3) = 0$$

and

$$\Sigma(vW_4) = 0$$

In summary: When the method of least squares is employed in determining the most probable values of certain unknown quantities, having as known coefficients the quantities W_1, W_2, W_3, W_4, the following relations hold as a necessary condition of the least squares method:

$$\Sigma(vW_1) = 0$$
$$\Sigma(vW_2) = 0$$
$$\Sigma(vW_3) = 0$$
$$\Sigma(vW_4) = 0$$

A knowledge of these relationships gives us a method of securing readily the value $\Sigma(v^2)$ and the standard error of estimate. Assume that, by the method of least squares, we have determined the constants in an equation of the type

$$Y_c = aW_1 + bW_2 + cW_3 + dW_4$$

For each residual we have the relation

$$v = aW_1 + bW_2 + cW_3 + dW_4 - Y \qquad (1)$$

Multiplying throughout by v, and summing, we have

$$\Sigma(v^2) = a\Sigma(vW_1) + b\Sigma(vW_2) + c\Sigma(vW_3) + d\Sigma(vW_4) - \Sigma(Yv) \quad (2)$$

But

$$\Sigma(vW_1) = 0$$
$$\Sigma(vW_2) = 0$$
$$\Sigma(vW_3) = 0$$
$$\Sigma(vW_4) = 0$$

therefore,

$$\Sigma(v^2) = -\Sigma(Yv) \quad (3)$$

Multiplying each equation (1) throughout by Y, and adding, we have

$$\Sigma(Yv) = a\Sigma(W_1Y) + b\Sigma(W_2Y) + c\Sigma(W_3Y) + d\Sigma(W_4Y) - \Sigma(Y^2) \quad (4)$$

Substituting in (3) the equivalent of $\Sigma(Yv)$ we have

$$\Sigma(v^2) = \Sigma(Y)^2 - a\Sigma(W_1Y) - b\Sigma(W_2Y) - c\Sigma(W_3Y) - d\Sigma(W_4Y) - \ldots \quad (5)$$

This gives us a means of obtaining the value $\Sigma(v^2)$ without computing the separate residuals, a method which is applicable whenever the equation of the curve to be fitted is of the form, or may be reduced, by the use of logarithms, reciprocals or other manipulation, to the form

$$Y = aW_1 + bW_2 + cW_3 + dW_4$$

In applying this to a particular case it is necessary only to replace W_1, W_2, W_3, W_4, etc., by the functions which actually appear as coefficients of the unknown quantities in the original equation. Thus in fitting a curve the equation to which is

$$Y = a + bX + cX^2 + dX^3$$

we find, as noted above, that

$$W_1 = 1$$
$$W_2 = X$$
$$W_3 = X^2$$
$$W_4 = X^3$$

Making these substitutions in equation (5) above we have

$$\Sigma(v^2) = \Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY) - c\Sigma(X^2Y) - d\Sigma(X^3Y) \quad (6)$$

The standard error, S_y, is derived from the equation

$$S_y = \frac{\Sigma(d^2)}{N}$$ [1]

where d is used to represent a deviation from a fitted curve. The deviation, d, then, is but another term for the residual v. Accordingly, as a general expression for the standard error of Y, with W_1, W_2, W_3, and W_4 as independent variables, we have

$$S_y^2 = \frac{\Sigma Y^2 - a\Sigma(W_1Y) - b\Sigma(W_2Y) - c\Sigma(W_3Y) - d\Sigma(W_4Y)}{N} \quad (7)$$

As in the previous case, this may be applied to a particular problem by replacing W_1, W_2, W_3, W_4, etc., by the actual coefficients of the unknown quantities.

Derivation of the Formula for the Index of Correlation

We have adopted as an index of the degree of correlation between two variables the measure ρ (Rho), derived from the equation

$$\rho^2_{yx} = 1 - \frac{S_y^2}{\sigma_y^2} \quad (8)$$

assuming a single dependent variable, Y, and a single independent variable, X. With a single dependent variable, Y, and a number of independent variables, W_1, W_2, W_3, W_4, the expression might be written

[1] Since our object is to measure the actual "scatter" about the fitted curve, the formula $\frac{\Sigma(d^2)}{N}$ is used, rather than the formula $\frac{\Sigma(d^2)}{N - N_c}$ (where N represents the number of observations and N_c the number of constants in the equation to the fitted curve). The second formula would be used, in accordance with the theory of least squares, if we were seeking to determine the mean square error of an observation or of an observational equation.

$$\rho^2_{y \cdot w_1 w_2 w_3 w_4} = 1 - \frac{S_y^2}{\sigma_y^2} \tag{9}$$

Corresponding changes would be made in the subscripts for other changes in the symbols employed. The expression above is equivalent to

$$\rho^2_{y \cdot w_1 w_2 w_3 w_4} = 1 - \frac{\Sigma(d^2)}{\Sigma(y^2)}$$

where y represents a deviation from an origin at the mean of the Y's. But

$$\frac{\Sigma(y^2)}{N} = \frac{\Sigma(Y^2)}{N} - c_y^2$$

where Y represents the original values of the Y-variable and c_y represents the difference between the original origin and the mean of the Y's. (The symbols c_y and c_x should not be confused with c, one of the constants in the equation to the fitted curve.)

Accordingly, we have

$$\rho^2_{y \cdot w_1 w_2 w_3 w_4} = 1 - \frac{\Sigma(d^2)}{\Sigma(Y^2) - Nc_y^2} \tag{10}$$

But we have secured an expression for $\Sigma(v^2)$ (the equivalent of $\Sigma(d^2)$) which holds in the case of a curve fitted by the method of least squares. Taking first the general case, with a number of independent variables, and substituting the equivalent of $\Sigma(d^2)$ in the above equation, we have

$$\rho^2_{y \cdot w_1 w_2 w_3 w_4} \ldots = 1 - \frac{\Sigma(Y^2) - a\Sigma(W_1 Y) - b\Sigma(W_2 Y) - c\Sigma(W_3 Y) - d\Sigma(W_4 Y) - \ldots}{\Sigma(Y^2) - Nc_y^2}$$

Simplifying this, we have, as a general formula for the index of correlation

$$\rho^2_{y \cdot w_1 w_2 w_3 w_4} \ldots = \frac{a\Sigma(W_1 Y) + b\Sigma(W_2 Y) + c\Sigma(W_3 Y) + d\Sigma(W_4 Y) + \ldots - Nc_y^2}{\Sigma(Y^2) - Nc_y^2} \tag{11}$$

This may be applied to a specific case by replacing W_1, W_2, W_3, W_4, etc., in the above formula by the functions which appear as coefficients of the unknown quantities in the original equation. When all these are functions of a single independent variable, as in the usual case, the index of correlation would be represented by the symbol ρ_{yx}.

Certain Special Cases

In the case of multiple correlation, where the symbols X_1, X_2, X_3, X_4, etc., are used to represent all the variables, whether considered dependent or independent, the symbol R is employed for the measure of correlation and numerical subscripts utilized as described in the body of the text.

In the case of a straight line relationship between two variables, ρ is replaced by the symbol r, which represents the ordinary coefficient of correlation. As the general equation for r we have

$$r^2 = \frac{a\Sigma(Y) + b\Sigma(XY) - Nc_y^2}{\Sigma(Y^2) - Nc_y^2}$$

There are two special cases in which this formula may be simplified. If the origin be at the mean of the X's, we have

$$a = c_y = \frac{\Sigma Y}{N}$$

$$a^2 = c_y^2 = \frac{a\Sigma Y}{N}$$

$$Nc_y^2 = a\Sigma Y$$

and the formula for r reduces to

$$r^2 = \frac{b\Sigma(xY)}{\Sigma(Y^2) - Nc_y^2}$$

If the origin be at the mean of the Y's (it is not essential that it be also at the mean of the X's)

$$\Sigma(y) = 0, \text{ and } c_y = 0$$

and the formula for the coefficient of correlation becomes

$$r^2 = \frac{b\Sigma(Xy)}{\Sigma(y^2)}$$

In this latter case the general formula for ρ may also be simplified by the elimination of the terms $a\Sigma(y)$ and Nc_y^2.

Checks on the Formation of the Normal Equations

There are so many possibilities of arithmetical error in the formation and solution of a set of normal equations that checks should be employed wherever possible. A convenient check on the calculations leading to the normal equations is afforded by the introduction in each observation equation of an additional term, s, equal to the sum of all the known quantities in that equation. Thus, in the following system of observation equations, formed in fitting a line to the points 1, 3; 2, 4; 3, 6; 4, 5; 5, 10; 6, 9; 7, 10; 8, 12; 9, 11, the values of s are as indicated:

	s
$3 = a + 1b$	5
$4 = a + 2b$	7
$6 = a + 3b$	10
$5 = a + 4b$	10
$10 = a + 5b$	16
$9 = a + 6b$	16
$10 = a + 7b$	18
$12 = a + 8b$	21
$11 = a + 9b$	21

(The coefficient of a in each case is 1, and this is added to the other known quantities.)

In fitting a curve described by the type equation

$$Y = aW_1 + bW_2 + cW_3 + dW_4$$

the following relations prevail between s and the other quantities computed. For each observation equation,

$$Y + W_1 + W_2 + W_3 + W_4 = s$$

For the normal equations,

$$\Sigma(W_1Y) + \Sigma(W_1^2) + \Sigma(W_1W_2) + \Sigma(W_1W_3) + \Sigma(W_1W_4) = \Sigma(W_1s)$$
$$\Sigma(W_2Y) + \Sigma(W_1W_2) + \Sigma(W_2^2) + \Sigma(W_2W_3) + \Sigma(W_2W_4) = \Sigma(W_2s)$$
$$\Sigma(W_3Y) + \Sigma(W_1W_3) + \Sigma(W_2W_3) + \Sigma(W_3^2) + \Sigma(W_3W_4) = \Sigma(W_3s)$$
$$\Sigma(W_4Y) + \Sigma(W_1W_4) + \Sigma(W_2W_4) + \Sigma(W_3W_4) + \Sigma(W_4^2) = \Sigma(W_4s)$$

This form is capable of application to any specific problem. In each case the s-equations are formed in precisely the same way as the corresponding normal equations.

In applying these checks several additional columns are needed in the working tables, but the extra trouble is more than compensated by the opportunity to check the work at each stage. The application is illustrated in the following working table, showing the calculations involved in fitting a second degree parabola of the form

$$Y = a + bX + cX^2$$

to the nine points 1, 2; 2, 6; 3, 7; 4, 8; 5, 10; 6, 11; 7, 11; 8, 10; 9, 9.

TABLE A

Illustrating the Use of Checks on the Formation of Normal Equations

Y	X	X^2	XY	X^2Y	s	Xs	X^2s
2	1	1	2	2	5	5	5
6	2	4	12	24	13	26	52
7	3	9	21	63	20	60	180
8	4	16	32	128	29	116	464
10	5	25	50	250	41	205	1,025
11	6	36	66	396	54	324	1,944
11	7	49	77	539	68	476	3,332
10	8	64	80	640	83	664	5,312
9	9	81	81	729	100	900	8,100
74	45	285	421	2,771	413	2,776	20,414

(Columns for X^3 and X^4 are omitted, as the values $\Sigma(X^3)$ and $\Sigma(X^4)$ may be derived from prepared tables.)

Each of the values in the column headed s is secured from the corresponding observation equation. Thus, from the first observation equation

$$2 = 1a + 1b + 1c$$

we have 5 as the value of s (2, plus the coefficients of the three constants). These values of s are secured readily from the table by adding the figures in the columns headed Y, X and X^2, plus 1, the coefficient of the constant term a.

Adding the various columns, the arithmetic work is verified by the following checks:

$$\Sigma(Y) + N + \Sigma(X) + \Sigma(X^2) = \Sigma(s)$$
$$74 + 9 + 45 + 285 = 413$$
$$\Sigma(XY) + \Sigma(X) + \Sigma(X^2) + \Sigma(X^3) = \Sigma(Xs)$$
$$421 + 45 + 285 + 2{,}025 = 2{,}776$$
$$\Sigma(X^2Y) + \Sigma(X^2) + \Sigma(X^3) + \Sigma(X^4) = \Sigma(X^2s)$$
$$2{,}771 + 285 + 2{,}025 + 15{,}333 = 20{,}414$$

Further uses of a check of this kind are explained below, in discussing the solution of the normal equations.

Other Tests

The possibility of checking the calculations in other ways has been suggested in the preceding sections. Thus, where the coefficients of the constants in the equation to the fitted curve are represented by W_1, W_2, W_3, W_4, we know that

$$\Sigma(vW_1) = 0$$
$$\Sigma(vW_2) = 0$$
$$\Sigma(vW_3) = 0$$
$$\Sigma(vW_4) = 0$$

If a curve of the type

$$Y = a + bX + cX^2 + dX^3$$

has been fitted, this means that

$$\Sigma(v) = 0$$
$$\Sigma(vX) = 0$$
$$\Sigma(vX^2) = 0$$
$$\Sigma(vX^3) = 0$$

The accuracy of the work may be tested by checking these relations.

Finally, we may test the accuracy of the work by computing the standard error of estimate in two different ways. We may compute the separate residuals by taking the difference between computed and actual values of the dependent variable, and from these values determine S. This may be compared with the results secured by applying the general formula for the standard error, as derived above. In the fitting of the second degree parabola, the data of which were used to illustrate the method of checking the normal equations, the equation derived was

$$Y = -.92860 + 3.52316X - .267316X^2$$

From the residuals separately computed we have

$$S_y = .4941$$

From the formula

$$S_y^2 = \frac{\Sigma(Y^2) - a\Sigma(Y) - b\Sigma(XY) - c\Sigma(X^2Y)}{N}$$

we have

$$S_y = .4947$$

This constitutes a final check upon the accuracy of the calculations.

Solution of the Normal Equations

The task of solving the normal equations is not a difficult one in most of the cases presented to the economic statistician. If there are only two or three unknowns the corresponding number of normal equations may be solved by simple algebraic methods. Even with three equations, however, it is advisable to employ a systematic procedure, and with more than three equations this is imperative. Such systematic methods of solving the simultaneous equations which are met with in connection with the method of least squares have been worked out by Gauss and by

Doolittle. The latter method, which is perhaps the more convenient for general usage, is demonstrated below.

The coefficients of the unknowns in the normal equations are always symmetrical with respect to the principal diagonal. Thus in securing the most probable values of the constants in the equation

$$Y = aW_1 + bW_2 + cW_3 + dW_4$$

we have the four normal equations

$$\begin{aligned}
a\Sigma(W_1^2) + b\Sigma(W_1W_2) + c\Sigma(W_1W_3) + d\Sigma(W_1W_4) - \Sigma(W_1Y) &= 0\\
a\Sigma(W_1W_2) + b\Sigma(W_2^2) + c\Sigma(W_2W_3) + d\Sigma(W_2W_4) - \Sigma(W_2Y) &= 0\\
a\Sigma(W_1W_3) + b\Sigma(W_2W_3) + c\Sigma(W_3^2) + d\Sigma(W_3W_4) - \Sigma(W_3Y) &= 0\\
a\Sigma(W_1W_4) + b\Sigma(W_2W_4) + c\Sigma(W_3W_4) + d\Sigma(W_4^2) - \Sigma(W_4Y) &= 0
\end{aligned}$$

The symmetrical arrangement about the diagonal, when Y-terms are neglected, is obvious. Starting with any term on the principal diagonal, we have the same coefficients directly above as to the left. Thus above the diagonal term in which the coefficient $\Sigma(W_3^2)$ appears we have the coefficients $\Sigma(W_2W_3)$ and $\Sigma(W_1W_3)$. The same coefficients are found to the left of the given diagonal term, and on the same line. The terms to the left of each diagonal entry may be omitted, therefore, and we may write the normal equations in the form

$$\begin{aligned}
a\Sigma(W_1^2) + b\Sigma(W_1W_2) + c\Sigma(W_1W_3) + d\Sigma(W_1W_4) - \Sigma(W_1Y) &= 0\\
+ b\Sigma(W_2^2) + c\Sigma(W_2W_3) + d\Sigma(W_2W_4) - \Sigma(W_2Y) &= 0\\
+ c\Sigma(W_3^2) + d\Sigma(W_3W_4) - \Sigma(W_3Y) &= 0\\
+ d\Sigma(W_4^2) - \Sigma(W_4Y) &= 0
\end{aligned}$$

The Doolittle Method

The Doolittle method may be illustrated with reference to the normal equations given on page 494 of the text. In the abbreviated form given above these are

$$\begin{aligned}
8.3564b_{12.34} + 2.790b_{13.24} + 2.932b_{14.23} + 47.967 &= 0\\
+ 6.6645b_{13.24} + 2.063b_{14.23} + 62.039 &= 0\\
+ 7.7893b_{14.23} + 47.519 &= 0
\end{aligned}$$

We wish to solve these for the constants. For convenience of reference let us represent the constant $b_{12.34}$ by A, $b_{13.24}$ by B, and $b_{14.23}$ by C. All the work of computation, with the necessary checks, is shown in the following table:

TABLE B

Solution of Normal Equations by the Doolittle Method

Line	(1) Reciprocals	(2) A	(3) B	(4) C	(5)	(6) s	
I		8.3564	2.790	2.932	47.967	62.0454	
II			6.6645	2.063	62.039	73.5565	
III				7.7893	47.519	60.3033	
1		8.3564	2.790	2.932	47.967	62.0454	
2	−.11966876	−1.000000	−.333876	−.350869	−5.740151	−7.424896	check
3			6.6645	2.063	62.039	73.5565	
4			−.931514	−.978924	−16.015030	−20.715470	
5			5.732986	1.084076	46.023970	52.841030	check
6	−.17442917		−1.000000	−.189094	−8.027923	−9.217017	check
7				7.7893	47.519	60.3033	
8				−1.028748	−16.830133	−21.769807	
9				−.204992	−8.702857	−9.991922	
10				6.555560	21.986010	28.541571	check
11	−.15254227			−1.000000	−3.353796	−4.353796	check

Back Solution

C	B	A
−3.353796	−8.027923	−5.740151
−3.353796	+.634183	+2.468592
	−7.393740	+1.176743
		−2.094816

$$A = b_{12.34} = -2.094816$$
$$B = b_{13.24} = -7.393740$$
$$C = b_{14.23} = -3.353796$$

Check:

Equation I:

$$8.3564b_{12.34} + 2.790b_{13.24} + 2.932b_{14.23} = -47.967$$

Substituting the given values,

$$8.3564\,(-2.094816) + 2.790\,(-7.393740) + 2.932\,(-3.353796) = -47.966985$$

Explanation. — The coefficients of the unknown quantities, A, B, and C, are listed in the designated columns. The known term in each normal equation is listed in column (5). (The sign of this known term, it should be noted, is that which it would have when the entire expression, of which it is one term, is equated to zero.) Column s is employed as a check. The value in column s, in each of the lines I, II and III, is the algebraic sum of the known values in the given normal equation. In securing this sum the coefficients to the left of the diagonal, which have been omitted from the table as it stands, must be included.

The following is a summary of the procedure in solving the normal equations:

1. In line (1) write normal equation I.

2. In line (2), column (1), write the reciprocal of the value in line (1), column (2), *with sign changed.* (This is the reciprocal of the coefficient of A.) Multiply each item in line (1) by this reciprocal, entering the products in the corresponding columns in line (2). (The algebraic sum of the items in columns (2), (3), (4) and (5) of line (2) should equal the value in column (6).) This operation has eliminated the unknown A, by expressing it in terms of B and C. (The -1 in line (2), column (2), has been included only to facilitate the checking process. The same is true in lines (6) and (11).) A heavy line may be drawn across the table below line (2).

3. Write normal equation II in line (3).

4. Multiply by the coefficient of B in line (2) (i.e., $-.333876$) the items in columns (3), (4), (5) and (6) in line (1). Enter the products in the corresponding columns of line (4).

5. Add lines (3) and (4), entering the sums in line (5). (The algebraic sum of the items in columns (3), (4) and (5) of line (5) should equal the value in column (6).)

6. In column (1), line (6), enter the reciprocal of the value in column (3), line (5), *reversing the sign.* Multiply each term in line (5) by this reciprocal, entering the products in line (6). (The sum of the items in columns (3), (4) and (5) of line (6)

should equal the value in column (6).) This operation has eliminated the unknown B, by expressing it in terms of C. A heavy line may be drawn across the table below line (6).

7. Write normal equation III in line (7).

8. Multiply by the coefficient of C in line (2) (i.e., $-.350869$) the items in columns (4), (5), and (6) of line (1). Enter the products in the corresponding columns of line (8).

9. Multiply by the coefficient of C in line (6) (i.e., $-.189094$) the items in columns (4), (5) and (6) of line (5). Enter the products in the corresponding columns of line (9).

10. Add lines (7), (8) and (9), entering the sums in line (10). (The algebraic sum of the items in columns (4) and (5) of line (10) should equal the value in column (6).)

11. In column (1), line (11), enter the reciprocal of the value in column (4) of line (10), *reversing the sign*. Multiply each term in line (10) by this reciprocal, entering the products in line (11). (The algebraic sum of the items in columns (4) and (5) of line (11) should equal the value in column (6).) This operation gives the value of C, which is found in column (5) of line (11). A heavy line may be drawn across the table below line (11).

Were there additional unknowns, as D and E, this last operation would have given C as a function of D and E and it would be necessary to carry the process still further, repeating the steps taken above. The next operation would be to bring down the fourth normal equation, entering it in line (12). Then the coefficients of D in lines (2), (6) and (11) would be used to multiply the necessary items in lines (1), (5) and (10), the products being entered in lines (13), (14) and (15). The sum of the items in lines (12), (13), (14) and (15) would be entered in line (16) and checked by the item in the s column. Multiplying through by the reciprocal of the coefficient of D in line (16), with sign reversed, the value of D would be obtained in terms of E. The value of E would be derived in a similar fashion.

The checks on these various operations have been indi-

cated in the table. The testing of the results at each step reduces the possibility of error to a minimum.

The back solution presents no difficulties. We have, from line (11),

$$C = -3.353796$$

from line (6)

$$B = -.189094C - 8.027922$$

from line (2)

$$A = -.333876B - .350869C - 5.740151$$

(The items in column (6) are inserted merely as checks. The items -1.000000 which appear in lines (2), (6) and (11) are inserted to assist in the checking.)

The computations involved in the back solution appear in the table.

A final check is afforded by inserting the values secured by this process in one of the normal equations. This check, as carried out for equation I, is shown below the table.

REFERENCES

Adams, Oscar S. *Application of the Theory of Least Squares to the Adjustment of Triangulation.* Special Publication No. 28, U. S. Coast and Geodetic Survey, 1915.

Brunt, David. *The Combination of Observations.*

Huntington E. V. *Curve Fitting by the Method of Least Squares and the Method of Moments* (in Rietz, H. L. ed. *Handbook of Mathematical Statistics*, 62–70).

Merriman, Mansfield. *The Method of Least Squares.*

Smith, Bradford B. *The Use of Punched Card Tabulating Equipment in Multiple Correlation Problems.* Washington, Bureau of Agricultural Economics, 1923.

Weld, L. D. *Theory of Errors and Least Squares.*

Whitaker, E. T. and Robinson, G. *The Calculus of Observations* (209–259).

Wright and Hayford. *Adjustment of Observations.*

APPENDIX B

GLOSSARY OF SYMBOLS

The following are the most important symbols employed in the preceding pages. A given symbol is sometimes called upon to serve different purposes, but the precise meaning should be clear from the context.

1. General symbols for variables and constants:

x: a variable quantity.

y: a variable quantity.

In general, any letter near the end of the alphabet may be employed to represent a variable quantity. Different variable quantities may be represented by the use of a single symbol, with different subscripts, as X_1, X_2, X_3, or W_1, W_2, W_3. [A distinction is later drawn (*cf.* Symbols employed in the measurement of relationship) between capital letters and small letters, as used to represent variable quantities.]

a: a constant (i.e., a quantity the value of which does not change in the given discussion). In general, any letter near the beginning of the alphabet may be used to represent a constant.

2. Symbols employed in the analysis and description of the frequency distribution:

m: the value of an individual observation; the value of the mid-point of a class. (The symbols a_1, a_2, a_3 are sometimes employed to represent different observations in a series.)

f: the number of observations in a given class; the frequency of a given class.

i: the class-interval.

l: the lower limit of a class.

N: the total number of cases in a given series or frequency distribution.

d: the deviation of a given observation from an average; usually, a deviation from the arithmetic mean. When written with a subscript, as d_x or d_y, it refers to a deviation from the arithmetic mean of the variable represented by the subscript. The symbol d is sometimes used to designate the difference between mean and mode.

d': the deviation of a given observation from an arbitrary origin, or assumed mean.

c: the difference between an arbitrary origin, or assumed mean, and the true mean (in terms of the symbols explained below, $c = M - M'$).

Σ(Sigma): the symbol for the process of summation. Thus Σd means the sum of all the deviations.

w_1, w_2, w_3: weights attached to a series of measures being averaged. (Not to be confused with similar symbols used to represent different variable quantities.)

y_o: the maximum ordinate of a frequency curve.

Symbols for averages, quartiles, etc.:

M: the arithmetic mean.

$M.d.$: the median.

$M.o.$: the mode.

M_g: the geometric mean.

H: the harmonic mean.

M': the value of an assumed arithmetic mean.

Q_1: the first or lower quartile.

Q_2: the second quartile or median.

Q_3: the third or upper quartile.

K: the value of a point midway between the first and third quartiles

D_3: the third decile.

Symbols for measures of variation and skewness:

$M.D.$: the mean deviation.

σ: the standard deviation; the root-mean-square deviation about the arithmetic mean.

s: the root-mean-square deviation about an origin other than the arithmetic mean.

$P.E.$: the probable error.
$Q.D.$: the quartile deviation.
q_1: the difference between the median and the lower quartile $(M.d. - Q_1)$.
q_2: the difference between the upper quartile and the median $(Q_3 - M.d.)$.
V: the coefficient of variation.
sk: a measure of skewness.
χ(Chi): a measure of skewness based upon the criteria β_1 and β_2. (These criteria are explained below.)

3. Symbols relating to index numbers.

p_o': price of a given commodity at time "0" (the base period).
q_o': quantity of same commodity at time "0".
p_1': price of same commodity at time "1".
q_1': quantity of same commodity at time "1".
p_o'': price of a second commodity at time "0".
q_o'': quantity of second commodity at time "0".
p_1'': price of second commodity at time "1".
q_1'': quantity of second commodity at time "1".
$\frac{p_1'}{p_o'}$: a price relative (relation of price of a given commodity at time "1" to price of same commodity at time "0").
$\frac{q_1'}{q_o'}$: a quantity relative.
P_o: price level at time "0".
P_1: price level at time "1".

4. Symbols employed in the measurement of relationship.

X: an observed value of a variable quantity.
Y: an observed value of a variable quantity. (The observed values of different variables may be represented also by the symbols X_1, X_2, X_3, or W_1, W_2, W_3).
$\overline{X}$: the arithmetic mean of a number of observed values of the variable X. A similar symbol may be employed for other variables. (In one demonstration in the preceding pages, relating to multiple correlation, the symbols A_1, A_2, A_3 . . . are used to represent the arithmetic means

of the variables X_1, X_2, X_3 . . . The symbols M_x and M_y are occasionally employed to designate the arithmetic means of different variables.)

x: value of a variable quantity expressed as a deviation from the arithmetic mean of all the observed values. The symbol y and the symbols x_1, x_2, x_3 . . . are similarly employed with respect to variables represented, as to original observations, by the symbols Y, X_1, X_2, X_3 . . .

x': a value of a variable quantity expressed as a deviation from an arbitrary origin. The symbol y' has a similar meaning.

Y_c: the computed or estimated value of a variable, as determined from an equation of average relationship; the symbol y_c may be employed for such a computed value, expressed as a deviation from the mean.

p: the mean product of two variables when expressed as deviations from their respective arithmetic means, i.e., $p = \frac{\Sigma(xy)}{N}$. When written with subscripts, as p_{12}, the latter relate to the variables in question, as x_1, x_2.

p': the mean product of two variables when expressed as deviations from assumed arithmetic means, i.e., $p' = \frac{\Sigma(x'y')}{N}$.

r: the Pearsonian coefficient of correlation. When written with subscripts, the latter indicate the variables to which the coefficient relates. Thus r_{yx} refers to the variables y and x, and r_{12} refers to the variables x_1 and x_2.

ρ(Rho): a general index of correlation (not to be confused with Spearman's coefficient of correlation based upon the squares of differences in rank). Subscripts should be employed to indicate the variables to which the measure relates, as ρ_{yx}, ρ_{xy}, $\rho_{\log yx}$, $\rho_{\log y \log x}$, $\rho_{\frac{1}{y}x}$, etc. (In each case the first subscript relates to the dependent variable.)

d: the deviation of a given observation from a fitted curve; the difference between an observed and a corresponding computed value of a variable.

v: a residual; identical in meaning with d, as given above.

S: the root-mean-square deviation about a fitted curve; the standard error of estimate. This measure should be written with a subscript to indicate the variable to which it applies, as S_y, S_x, $S_{\log y}$ (the standard error of estimate in terms of logarithms), S_r (the standard error of estimate in terms of ratios), $S_{\frac{1}{y}}$ (the standard error of estimate in terms of reciprocals).

η (Eta): the correlation ratio. Subscripts should be employed to represent the variables to which the measure relates, as η_{yx} or η_{xy}. The first subscript in each case relates to the dependent variable.

σ_{ay}: the root-mean-square deviation about a line through the means of the columns of a correlation table; the standard deviation of the y-arrays about their respective means. The symbol σ_{ax} has the same meaning with respect to the rows of a correlation table, or the x-arrays.

σ_{my}: the standard deviation of the means of the columns of a correlation table about the mean of all the y's, the mean of each column being weighted by the number of items in that column. The symbol σ_{mx} has the same meaning with respect to the means of the rows.

ζ (Zeta): the test for linearity of regression ($\zeta = \eta^2 - r^2$).

κ: the number of arrays employed in the computation of a given correlation ratio.

b: the coefficient of regression; the slope of a line of regression. When written with subscripts, the latter relate to the variables in question, as b_{yx}, b_{12} (for the variables x_1, x_2). The first subscript relates to the dependent variable in each case; b_{yx} is the coefficient of regression of y on x and b_{xy} is the coefficient of regression of x on y.

$R_{1.234}$: the coefficient of multiple correlation between a dependent variable, x_1, and a combination of independent variables, x_2, x_3, and x_4. The order may be changed, but the primary subscript always relates to the dependent variable.

$r_{12.34}$: the coefficient of partial or net correlation between the variables x_1 and x_2, when the variables x_3 and x_4 are held constant. The order of subscripts is changed for a different combination of variables, the two primary

subscripts always relating to the variables between which the net correlation is being measured.

$b_{12.34}$: the coefficient of net regression between the variables x_1 and x_2, the former being dependent, when the variables x_3 and x_4 are also taken account of in the estimating equation; the weight given to x_2 in estimating x_1, when the estimate is also based upon values of x_3 and x_4. The order of subscripts is changed for a different combination of variables.

$S_{1.234}$: the root-mean-square deviation about a line describing the relationship between a dependent variable, x_1, and a series of independent variables, x_2, x_3 and x_4; the standard error of estimate of x_1 under these conditions.

$\sigma_{1.234}$: the standard deviation of the fourth order; identical with $S_{1.234}$.

(In all the measures immediately above, the number of subscripts corresponds to the number of variables included in a given study. For the sake of simplicity, only four variables have been assumed.)

5. Symbols employed in the measurement of errors.

σ_M: the standard error of the arithmetic mean.

σ_σ: the standard error of the standard deviation. Similarly, the symbol σ, with any subscript, represents the standard error of the measure to which the subscript relates.

$P.E._M$: the probable error of the arithmetic mean ($P.E._M = .67449\ \sigma_M$). Similarly, the symbol $P.E.$ with any subscript represents the probable error of the measure to which the subscript relates.

6. Other symbols.

p: the probability of the successful outcome of a given event.

q: the probability of an unsuccessful outcome of a given event.

s: the sum of the known quantities in an observation equation; used in checking the formation and solution of the normal equations.

Δ (Delta): a general symbol for the difference between two magnitudes or between successive values of a variable

quantity. The variable to which it refers may be represented by a subscript, as Δ_y. First differences are represented by Δ^1, second differences by Δ^2, etc.

ν_1, ν_2, ν_3, etc.: moments of a frequency distribution about an arbitrary origin.

π_1, π_2, π_3, etc.: uncorrected moments of a frequency distribution about the arithmetic mean.

μ_1, μ_2, μ_3, etc.: moments of a frequency distribution about the arithmetic mean after the application of Sheppard's corrections. (If Sheppard's corrections are not necessary, these symbols would be employed for the uncorrected moments about the mean.)

β_1: $\dfrac{\mu_3^2}{\mu_2^3}$

β_2: $\dfrac{\mu_4}{\mu_2^2}$

κ_2: A criterion of curve type based on β_1 and β_2.

χ^2: A quantity used in testing goodness of fit.

LIST OF REFERENCES

ADAMS, OSCAR S. *Application of the Theory of Least Squares to the Adjustment of Triangulation.* Special Publication No. 28. U. S. Coast and Geodetic Survey. 1915.

BARLOW. *Tables of Squares, Cubes, Square Roots, Cube Roots, Reciprocals.* Spon and Chamberlain. New York. 1919.

BECKETT, S. H., and ROBERTSON, R. D. *The Economical Irrigation of Alfalfa in the Sacramento Valley.* Bulletin 280, Agricultural Experiment Station. University of California. May, 1917.

BOWLEY, A. L. *Elements of Statistics.* P. S. King and Son, London. 1920.

BRINTON, W. C. *Graphic Methods for Presenting Facts.* The Engineering Magazine Co., New York. 1914.

BROAD, C. D. *On the Relation between Induction and Probability.* Mind N. S. Vol. 27, 1918, Vol. 29, 1920.

BRUNT, DAVID. *The Combination of Observations.* Cambridge University Press. 1917.

CHADDOCK, R. E. *Principles and Methods of Statistics.* Houghton Mifflin, Boston. (*In preparation.*)

CLARK, WALLACE. *The Gantt Chart.* Ronald Press, New York. 1922.

CRUM, W. S. *The Use of the Median in Determining Seasonal Variation.* Journal of the American Statistical Association. March, 1923.

DAVENPORT, E. *Comparative Agriculture.* In Bailey's Cyclopedia of American Agriculture.

DAVIES, G. R. *Introduction to Economic Statistics.* Century, New York. 1922.

DAY, EDMUND E. *An Index of the Physical Volume of Production.* Review of Economic Statistics. Sept., 1920, Jan., 1921. *Standardization of the Construction of Statistical Tables.* Quarterly Publications of the American Statistical Associ-

ation. March, 1920. *The Volume of Production of Basic Raw Materials in the United States.* Review of Economic Statistics. July, 1922.

EDGEWORTH, F. Y. *On Correlated Averages.* Phil. Mag., 5th series. Vol. 34. 1892.

EDITORIAL. *On the Probable Errors of Frequency Constants.* Biometrika, Vol. 2 (273–281).

ELDERTON, W. PALIN. *Frequency Curves and Correlation.* Layton, London. 1906.

EZEKIEL, M. J. B. *A Method of Handling Curvilinear Correlation for Any Number of Variables,* Journal of the American Statistical Association, Dec., 1924.

FAULKNER, HELEN D. *The Measurement of Seasonal Variation.* Journal of the American Statistical Association, June, 1924.

FIELD, J. H. *Some Advantages of the Logarithmic Scale in Statistical Diagrams.* Journal of Political Economy. Oct., 1917.

FISHER, ARNE. *An Elementary Treatise on Frequency Curves.* Macmillan, New York. 1922. *The Mathematical Theory of Probabilities.* Macmillan, New York. 1922.

FISHER, IRVING. *A Weekly Index Number of Wholesale Prices.* Journal of the American Statistical Association. Sept,, 1923. *The Making of Index Numbers.* Houghton Mifflin, Boston. 1922. *The "Ratio" Chart.* Quarterly Publications of the American Statistical Association. June, 1917. *Revision of the Weekly Index Number.* Journal of the American Statistical Association, Sept., 1924.

FLUX, A. W. *The Measurement of Price Changes.* Journal of the Royal Statistical Society. March, 1921.

GALTON, FRANCIS. *Correlations and Their Measurement.* Proceedings of the Royal Society. Vol. 45, 1888.

GRIFFIN, F. L. *Introduction To Mathematical Analysis.* Houghton Mifflin, Boston. 1922.

HAAS, G. C. *Sale Prices as a Basis for Farm Land Appraisal.* Technical Bulletin, No. 9. University of Minnesota Agricultural Experiment Station. Nov., 1922.

HALL, LINCOLN W. *Seasonal Variation as a Relative of Secular Trend.* Journal of the American Statistical Association, June, 1924.

HART, W. L. *The Method of Monthly Means for Determination*

of a Seasonal Variation. Journal of the American Statistical Association. Sept., 1922.

Haskell, A. C. *Graphic Charts in Business.* Codex Book Co., New York. 1922.

How to Make and Use Graphic Charts. Codex Book Co., New York. 1919.

Jones, D. C. *A First Course in Statistics.* Bell, London. 1921.

Karsten, Karl. *Charts and Graphs.* Prentice Hall, New York. 1923.

Kelley, Truman L. *Statistical Method.* Macmillan, New York. 1923.

Keynes, J. M. *A Treatise on Probability.* Macmillan, New York. 1921.

Killough, H. B. *A Statistical Analysis of Oat Prices.* Bureau of Agricultural Economics.

King, W. I. *Elements of Statistical Method.* Macmillan, New York. 1912. *An Improved Method for Measuring the Seasonal Factor.* Journal of the American Statistical Association, Sept., 1924.

Knibbs, G. H. *The Theory and Justification of Curve Smoothing.* In H. Secrist, *Readings and Problems in Statistical Methods.* Macmillan, New York. 1920.

Kurtz, Edwin. *Replacement Insurance.* Administration, July, 1921.

Lipka, Joseph. *Graphical and Mechanical Computation.* Wiley, New York. 1918.

Mellor, J. W. *Higher Mathematics for Students of Chemistry and Physics.* Longmans, London. 1922.

Merriman, Mansfield. *The Method of Least Squares.* Wiley, New York. 1897.

Mills, Frederick C. *On Measurement in Economics* (in Tugwell, R. G. ed., *The Trend of Economics.* Knopf, New York. 1924).

Miner, J. R. *Tables of* $\sqrt{1-r^2}$ *and* $1-r^2$ *for use in Partial Correlation and Trigonometry.* Johns Hopkins Press, Baltimore. 1922.

Mitchell, W. C. *History of Prices During the War.* Price Bulletin, No. 1, War Industries Board 1919. *The Making and Using of Index Numbers.* Part I. Bulletin No. 284. U. S. Bureau of Labor Statistics. Oct., 1921.

MOORE, H. L. *Economic Cycles: Their Law and Cause.* Macmillan, New York. 1914. *Elasticity of Demand and Flexibility of Prices.* Journal of the American Statistical Association. March, 1922. *Empirical Laws of Demand and Supply and the Flexibility of Prices.* Political Science Quarterly. Dec., 1919. *Forecasting the Yield and the Price of Cotton.* Macmillan, New York. 1917. *Generating Economic Cycles.* Macmillan, New York. 1923.

NATIONAL BUREAU OF ECONOMIC RESEARCH. *Income in the United States.* (Edited by W. C. Mitchell.) Harcourt Brace and Co., New York. 1921.

NORTH DAKOTA AGRICULTURAL COLLEGE. *Cost of Production and Farm Organization on 126 Farms in North Dakota.* Bulletin No. 165, Agricultural Experiment Station. 1922.

OGBURN, W. F., and THOMAS, DOROTHY. *Influence of the Business Cycle on Certain Social Conditions.* Quarterly Publications of the American Statistical Association. Sept., 1922.

PEAKE, E. G. *An Academic Study of Some Money Market and Other Statistics.* P. S. King, London. 1923.

PEARL, R., and REED, L. J. *Predicted Growth of Population of New York and Its Environs.* Committee of Plan of New York. 1923.

PEARL, RAYMOND. *Medical Biometry and Statistics.* Saunders, Philadelphia. 1923.

PEARSON, KARL. *Mathematical Contributions to the Theory of Evolution. On the General Theory of Skew Correlation and Non-Linear Regression.* Draper's Company Research Memoirs. Cambridge University Press. 1905. *Notes on the History of Correlation.* Biometrika, Vol. 13. *On a Correction Needful in the Case of the Correlation Ratio.* Biometrika, Vol. 8. *On the Correction Necessary for the Correlation Ratio.* Biometrika, Vol. 14. *Regression, Heredity and Panmixia.* Phil. transactions, Royal Society, Series A, Vol. 187. 1896. *Tables for Statisticians and Biometricians.* Cambridge University Press. 1914. *The Fundamental Problem of Practical Statistics.* Biometrika, Vol. 13.

PERSONS, WARREN M., and COYLE, EUNICE. *A Commodity Price Index of Business Cycles.* Review of Economic Statistics. Prel. Vol. 3.

PERSONS, WARREN M. *An Index of Trade for the United States.* Review of Economic Statistics. April, 1923. *Correlation of Time Series.* Journal of the American Statistical Association. June, 1923. *Fisher's Formula for Index Numbers.* Review of Economic Statistics. Prel. Vol. 3. *Indices of Business Conditions.* Review of Economic Statistics. Prel. Vol. 1. 1919. *The Variate Difference Correlation Method and Curve Fitting.* Quarterly Publications of the American Statistical Association. June, 1917.

PRESCOTT, RAYMOND. *Law of Growth in Forecasting Demand.* Journal of the American Statistical Association. Dec., 1922.

RIETZ, H. L. ed. *Handbook of Mathematical Statistics.* Houghton Mifflin, Boston. 1924.

RIETZ, H. L., and CRATHORNE, A. R. *College Algebra.* Holt, New York. 1917.

RUGG, H. O. *Statistical Methods Applied to Education.* Houghton Mifflin, Boston. 1917.

RUNNING, T. R. *Empirical Formulas.* Wiley, New York. 1917.

SCHULTZE, ARTHUR. *Graphic Algebra.* Macmillan, New York. 1918.

SCHULTZ, HENRY. *The Statistical Measurement of the Elasticity of Demand for Beef.* Journal of Farm Economics. July, 1924.

SECRIST, HORACE. *Introduction to Statistical Methods.* Macmillan, New York. 1917. *Readings and Problems in Statistical Methods.* Macmillan, New York. 1920.

SHEPPARD, W. F. *On the Calculation of the Most Probable Values of Frequency Constants for Data Arranged According to Equi-distant Divisions of a Scale.* Proceedings of the London Mathematical Society, Vol. 29. 1898. *The Calculation of the Moments of a Frequency Distribution.* Biometrika, Vol. 5.

SMITH, BRADFORD B. *The Use of Punched Card Tabulating Equipment in Multiple Correlation Problems.* (Prepared for the use of statisticians of the Bureau of Agricultural Economics. U. S. Dept. of Agriculture.) 1923.

SNOW, E. C. *Trade Forecasting and Prices.* Journal of the Royal Statistical Society. May, 1923.

SNYDER, CARL. *A New Index of the Volume of Trade.* Journal of the American Statistical Association. Dec., 1923.

STAMP, J. C. *The Effect of Trade Fluctuations upon Profits.* Journal of the Royal Statistical Society. July, 1918.

STEINMETZ, C. P. *Engineering Mathematics.* McGraw Hill, New York. 1917.

STEWART, ETHELBERT. *Labor Efficiency and Productiveness in Sawmills.* Monthly Labor Review. Jan., 1923.

TOLLEY, H. R., and EZEKIEL, M. J. B. *A Method of Handling Multiple Correlation Problems.* Journal of the American Statistical Association. Dec., 1923.

WALSH, C. M. *The Measurement of General Exchange Value.* Macmillan, New York. 1901. *The Problem of Estimation.* P. S. King and Son, London. 1921.

WELD, L. D. *Theory of Errors and Least Squares.* Macmillan, New York. 1916.

WEST, CARL J. *Introduction to Mathematical Statistics.* Adams, Columbus. 1918.

WHIPPLE, G. C. *Vital Statistics.* Wiley, New York. 1919.

WHITAKER, E. T., and ROBINSON, G. *The Calculus of Observations.* Blackie and Son, London. 1924.

WHITEHEAD, A. N. *An Introduction to Mathematics.* Holt, New York. 1911.

WORKING, HOLBROOK. *Factors Determining the Price of Potatoes in St. Paul and Minneapolis.* Technical Bulletin, No. 10, University of Minnesota Experiment Station. Oct., 1922.

WRIGHT, T. W., and HAYFORD, J. F. *Adjustment of Observations.* Van Nostrand, New York. 1906.

YOUNG, ALLYN A. *The Measurement of Changes in the General Price Level.* Quarterly Journal of Economics. Aug., 1921.

YULE, G. UDNEY. *An Introduction to the Theory of Statistics.* Griffin, London. 1919. *On the Time Correlation Problem, with Especial Reference to the Variate Difference Correlation Method.* Journal of the Royal Statistical Society. July, 1921.

ZIZEK, FRANZ. *Statistical Averages.* Holt, New York. 1913.

INDEX

Abscissa, 12
Acme Corporation, sales, 40, 41, 42
Administration, 3, 4
Aggregates of actual prices, simple, 193, 194, 195; weighted, 207, 208
Aggregative construction of index numbers of prices, 193
Agricultural products, 228
Agriculture. *See* U. S. Department of Agriculture
Alfalfa yield and irrigation, 433, 435, 436
Allocation of costs, 7
American index numbers of wholesale prices, 229; Bradstreet's, 234, 243; Dun's, 236, 243; Federal Reserve Board, 231, 243; Fisher's weekly index, 237, 243; Persons' commodity price index of business cycles, 238, 243; U. S. Bureau of Labor Statistics, 229, 242; Various Systems, 241; War Industries Board, 233, 243
American Telephone and Telegraph Co., 308, 309; composite index of general business activity, 357, 358; composite index correlated with stock price index, 422; general business curve, 356; study of annual message use of residence subscribers, 527, 528, 535, 536, 543
Analysis, 7
Animal and forest products, 228
Annalist's food price index, 241, 243
Anti-logarithms, 28
Arithmetic averages of relative prices, 195, 196; of egg prices, 316
Arithmetic mean, 112, 145, 453, 454; characteristic features, 144; computation of, 113; short method of computing, 115, 118; weighted, 115
Arithmetic measures, geometric and harmonic compared with, 475
Arithmetic series, 18
Array of data, 63, 64
Artillery firing, 100, 101, 102
Astronomical observations, 99
Asymmetry. *See* Skewness
Automobile speed, 141
Average relationship, 366, 367, 375
Averages, 97, 109, 111; arithmetic, 195, 196, 216; characteristic features of the chief averages, 144; criteria for use of, 140; in distribution of price changes, 187; references, 146; relations between different averages, 143; symbols, 583; weighted average, 115; when significant, 148. *See also* Mean; Moving averages

Bank clearings in New York City, measurement of trend, 259, 260, 261, 268
Bank discount rates, relations, 379, 380, 381, 382, 384, 385, 391, 395, 396, 397
Barlow's Tables, 142
Beckett, S. H., 433
Binomial expansion, 520
Bituminous coal production, 333; actual production, 337, 339, 340; changes in rate of growth, 335, 336; computation of cycles, 336, 337; corrected deviations from normal, together with actual monthly, 338; measurement of cyclical changes, 334
Blakeman, John, 557
Bonds, 125, 127
Bowley, A. L., 167
Bradley, 99
Bradstreet's index numbers of wholesale prices, 173, 195, 222, 226, 253, 307; description, 234, 243
Brinton, W. C., 51
Buffalo, N. Y., 527
Building contracts, 307; actual and de-

flated values, 313; A. T. & T. Co. deflated values, 308, 309
Business, 3; A. T. & T. Co. curve and composite index, 356, 357, 358; classes of activity, 3; quantitative character of its problems, 6
Business cycles, 8, 223; Persons' commodity price index, 238; stock price cycles as related to, 421, 423, 424, 425, 426
Business failures, 284; determining the secular trend, 284, 285, 288; second degree parabola and, 286; straight line and, 286; third degree parabola and, 287; with three lines of trend, 289

Catastrophes, 258
Central tendency, 107, 148; geometric mean as a measure of, 139; measures of, 109
Chain relatives, 319
Chaining index numbers, 218
Chance, 516, 518; comparison of actual and theoretical frequencies, 522. *See also* Probabilities
Characteristic, 27
Charlier, C. V. L., 157
Charts, 36; comparison of frequencies, 44; component parts, 45; construction, 36; cumulative, 46, 47; Gantt progress, 48, 49, 50, 51; logarithmic and semi-logarithmic, 32, 39; selection, 36; time series, 37
Chi-square test of goodness of fit of frequency curve, 543
Classes, 70; boundaries, 71; determination, 70; indeterminate, 72
Classification of quantitative data, 64
Class-intervals, 68; uniformity, 71
Class limits, 69
Coal. *See* Bituminous coal production
Coal and petroleum production, 171; index numbers, 170, 171; weighted index numbers, 172, 173
Coefficient of correlation, 372, 375, 391; coefficient of regression and, 503; effect of departure from normal, 406; measurement of time sequence and, 420; product-moment formula, 385; relation to correlation ratio, 450
Coefficient of multiple correlation, 496
Coefficient of regression, 394; coefficient of correlation and, 503
Coefficient of variation, 165
Coin tossing, 101, 102, 103
Column diagrams, 74, 75, 76, 77, 83, 84, 85
Commodities, change in prices, 104; number to be used in constructing index numbers, 222
Commodity index index of business cycles, 238
Comparability of data, 253
Comparison of frequencies, 44
Comparison of frequency distributions, 97
Component parts, 45
Compound interest, 34, 35; computing, 138
Concentration, 109; tendency toward, 111
Condensation of data, 7, 67
Constants, 14; symbols for, 14, 582
Continuous series, 85
Control, 5
Coördinate geometry, 11
Coördinates, 11, 13; on diagrams, 55; rectangular, 11
Corn yield and temperature, 485, 486, 510, 511; three independent variables and, 490
Correlation, linear, 363; multiple and partial, 485; non-linear, 432; partial and simple distinguished, 502; partial or net, 500; terms used in measurement, 393. *See also* Linear correlation; Multiple correlation, etc.
Correlation ratio, 441; computation, 442, 447; correction, 449; relation to coefficient of correlation, 450
Correlation table, 379; construction, 378; discount rates of Federal Reserve banks and of commercial banks, 381; tabulation of items, 380
Cost accounting, 7
Cost of living, index numbers, 245; wages in relation to, 250
Cotton, 13, 15; consumption, corrections for seasonal variation, 354; domestic consumption, index numbers, 169; production and prices, computing coefficient of correlation, 428; production and prices, correla-

tion of cyclical fluctuations, 413, 414, 415, 416
Coyle, E. S., 239
Cumulative arrangement of data, 88
Cumulative charts, 46
Cumulative frequency curve, 91, 92; weekly earnings of employees, 134
Cumulative tables, 90
Curve of error. *See* Normal curve of error
Curves, 16, 20; on diagrams, 55; factors determining choice, 290; hyperbolic, 20; normal, criteria, 531; parabolic, 20; potential series, fitting, 280; representation of secular trend by, 271; scale, reading, 56; sine, 20, 25, 26; smoothing, 80, 87; three families for relations of variables, 472; types, 20, 26. *See also* Fitting of curves
Cycles. *See* Business cycles
Cyclical fluctuations, 258; correlation between, 412; "lead" and "lag" in correlation, 420; measurement, 315, 332; references, 343

Data, comparability, 253; condensation, 67; cumulative arrangement, 88; homogeneity, 129; organization, 61; plotting, 300; quantitative, 6, 7; unorganized, 62
Dates on diagrams, 54
Davenport, D. H., 331, 461
Davis, Cal., 433
Day, E. E., 205, 347, 348, 350; 351
Deciles, 124; graphic location, 131, 133, 134
Deflating index, 308, 309 312
Deflation, 307
Depreciation, 90
Descartes, René, 11
Deviation, mean, 149, 150, 151; quartile, 158; standard, 154, 156, 162, 163, 368, 405, 513
Diagrams, 52
Dice-throwing experiment, 523, 554; actual and theoretical probabilities compared, 524, 525
Discount rates, 122
Discrete series, 85, 87
Dispersion, 147, 160. *See also* Variation
Dispersion, zone of, in artillery firing, 100
Dodge, F. W., & Co., 307
Doolittle method, 577, 578
Dun's index of wholesale prices, 236, 243

Economic data, 6; curves for, 297, 300; curves and histograms, 106; distribution, 103; plotting, 300
Economics, 3, 9; quantitative character of its problems, 6
Edgeworth, F. Y., 523
Egg prices, 315; arithmetic averages, 316; index numbers of seasonal variation based on percentage ratios to trend, 326; indices of seasonal variation, adjusted, 327; link relatives, 318, 319; moving averages, 321, 322
Elderton, W. P., 544
Elimination of irrelevant factors, 501
Empirical curves, 283, 284
Employees, distribution on the basis of weekly earnings, 65, 72, 74
Error, probable. *See* Probable error
Error, standard. *See* Standard error
Errors, symbols employed in measurement of, 587
Estimates, 370; assumptions involved, 453; problem of estimation, 453; references, 484; zones of estimate, 458; zones of estimate and their significance, 477
Exchange rates, 105, 106, 149, 151; London-Paris, 156, 164, 165
Expense items, 7
Exponent, 27
Exponential curve, 20, 23, 24
Extrapolation, 304
Ezekiel, M. J. B., 491, 499

Factor reversal test, 212, 213
Farm crop prices, 191; comparison of five types of index numbers, 201, 202; comparison of weighted index numbers, 216, 217; index numbers, 191, 192, 194; weighting of index numbers, 205
Farm products, index numbers of price and buying power, 246, 247
Farms in New England, 44, 45

Fechner, G. T., 125
Federal Reserve bank discount rates, relation to commercial bank rates, 379, 380, 381, 382, 384, 385, 391, 395, 396, 397
Federal Reserve Bank, New York, 223; deflation index, 312; index of trade volume, 359, 361; wholesale price index, 241, 243
Federal Reserve Board, index numbers of wholesale prices, 231, 243; index of production in basic industries, 352, 353; other index numbers, 355
First order coefficients of correlation, 507; computation, 509, 510
Fisher, Arne, 545
Fisher, Irving, factor reversal test, 212, 213; "ideal" index, 215, 217, 219, 346; new weekly index number of wholesale prices, 237, 243; substitute formula for "ideal" index, 218; time reversal test, 203, 204, 212; types of index numbers, 193; weighting methods for index numbers, 209
Fitting a straight line, 273; for nine points, 273; special cases, 278
Fitting of curves, 271; normal curve, 532, 536; potential series, 280; short cut methods, 298; testing goodness of fit of normal, 535, 540
Five per cent bonds, frequency distribution, 125
Fluctuations, 257; cyclical, measurement, 315; random, 258; seasonal, measurement, 315; short term, 427. *See also* Business cycles; Cyclical fluctuations
Food price indexes, 241; U. S. Bureau of Labor Statistics index of retail prices, 244
Frequencies, comparison, 44; determination of theoretical, 537, 540
Frequency curves, 45, 545; cumulative, 91; distribution of errors of astronomical observation, 99; distribution of soldiers by height, 98; ogive and, relation, 93, 94; personal income recipients, 86; smooth, 81; weekly earnings of employees, 132
Frequency distributions, 61; comparison, 97; description, 97, 147; description, note on, 545; employees on the basis of weekly earnings, 65, 72, 74; general characteristics, 107; graphic representation, 74; message use of residence (telephone) subscribers, illustrating computation of moments, 527, 528; methods of describing, 109; moments, 528; price ratios, 179; references, 96, 146, 168; symbols employed, 582
Frequency polygons, 77, 78, 79; change in commodity prices, 104, 180, 183, 186; changes plotted on logarithmic scale, 189; distribution of exchange rates, 105
Frequency series, 61, 92
Frequency tables, 64; construction, 67
Fulkner, H. D., 323
Functional relationship, 15

Galton, Francis, 394
Gantt, H. L., progress chart, 48
Gauss, K. F., 576
"General business" curve of the American Telephone and Telegraph Co., 356
Generalization, 550
Geometric averages of relative prices, 198
Geometric mean, 112; averaging ratios by, 188; characteristics, 136, 145; computing, 135; as a measure of central tendency, 139
Geometric measures, arithmetic and harmonic compared with, 475
Geometric series, 23
Geometry, coördinate, 11
Gibson, Thomas, food price index, 241
Gompertz curve, 297
Graphic presentation, 11, 36; frequency distributions, 74; references on methods, 59; standard rules, 51
Graphs, 36
Grouping process, 65
Growth, curve of, 297

Hall, L. W., 324
Harmonic averages of relative prices, 199
Harmonic mean, 112; characteristic features, 145; method of employing, 141

Harmonic measures, arithmetic and geometric compared with, 475
Height, human, 86, 87; soldiers', distribution, 98
Histogram, 76
Homogeneity of data, 129
Human heights, 86, 87, 98
Hyperbola, 22
Hyperbolic curve, 20

"Ideal" index, 215, 217, 219, 346; substitute formula for, 218
Income distribution, 82; frequency curve, 86
Index numbers, 169, 170; references, 251; symbols relating to, 193, 584; types and forms, 173, 174
Index numbers of physical volume, 344; references, 362
Index numbers of price and buying power of farm products, 246, 247
Index numbers of retail prices, 244
Index numbers of the cost of living, 245
Index numbers of wages, 248
Index numbers of wholesale prices, 169; commodities — number and kind to be used, 222, 223; comparison of simple types, 201, 202; differences exhibited by various systems, 242; fundamental types (Fisher's), 193; "ideal" index, 215, 217, 218 (formula), 219, 346; number of commodities to be included, 222; price groups in the field, 226; purpose, 177; reliability of different types, 220; simple aggregative type, 193; summary of characteristics and uses of different systems, 242; testing, 203, 212; time reversal test, 203, 204; various American systems described, 229; various methods employed in construction, 191; various problems involved, 221; weighted, comparison of farm crop prices, 216, 217; weighting, 205. *See also* American index numbers, etc.
Index of correlation, 436; based on logarithmic values, 462; derivation of the formula for, 570; meaning, 437; short method of computing, 438; standard error and, in terms of reciprocals, 468
Induction, 549; statistical, 548, 550
International index numbers of prices, 232
Interpolation, 81, 93
Irrigation, alfalfa yield and, 433, 435, 436

Joint Committee on Standards for Graphic Presentation, 51

Kansas, 486, 510
Karsten, K. G., on selection of curves, 305
Kelley, T. L., on the best type of index number, 220
Keynes, J. M., 552
Killough, H. B., 455, 456
Knibbs, G. H., 87
Kurtosis, 110, 147, 168, 545
Kurtz, Edwin, 88

Labor, "prices" of, 248
Least squares, method of, 273, 274; application to certain statistical problems, 562; for linear correlation, 375, 378, 381, 401; references, 581
Lettering of diagrams, 59
Life tables, 90
Linear correlation, 363; details of calculation, 375; least squares method of measurement, 375, 378, 381, 401; limitations of measures of relationship, 405; product-moment method of measurement, 385; references, 409
Linear magnitudes, 52
Linear relationship, 18
Lines of regression, 393, 394, 475
Link relatives, 318, 319; advantages, 328
Lipka, Joseph, 301, 303
Logarithmic charts, 32
Logarithmic coördinates on diagrams, 55
Logarithmic equations, 30
Logarithmic mean, 138
Logarithms, 26, 296; use in curve fitting, 290
London-New York Exchange rates, 105, 106, 149, 151
London-Paris exchange rates, 156, 164, 165

Macaulay, F. R., 352
Mantissa, 27
Manufactured goods, 227
Markets, 5
Mathematical concepts, 11; references, 59
Mathematical curves, representation of secular trend by, 271
Mathematical methods, 9, 11
Mean. *See* Arithmetic mean; Geometric mean; Harmonic mean
Mean deviation, 149, 161, 163; computation, 150, 151
Measurement, 8; methods, 9
Measurement of relationship, 363. *See also* Relationship
Measures, arithmetic, geometric, and harmonic, 474, 475
Measures of central tendency, 109, 111
Measures of reliability, 553; limitations, 559
Measures of skewness, 110, 166, 546
Measures of variation, 109, 147. *See also* Variation
Median, 112; characteristic features, 144; graphic location, 131, 132; location, 119, 121; of relative prices, 197
Mellor, J. W., 99
Merriman, Mansfield, 101
Mid-points, 69, 70, 71
Miner, J. R., 508
Mitchell, W. C., 104, 228, 231, 242; comparison of index numbers based on varying numbers of quotations, 222; on index numbers of prices, 181, 184; production index, 344
Modal group, 111
Mode, 111; characteristic features, 144; determining modal value from mean and median, 130; graphic location, 131, 132; location, 124
Moment, 528, 529
Monthly normal values, 329
Monthly values of production as basis for index numbers, 351
Mortality tables, 90, 522
Motor vehicle registration, relation to taxable personal incomes, 364, 365, 366, 369
Moving averages, 260; application, 263; application to series with linear trend, 264; application to non-linear series, 266, 267; characteristics, 262; measurement of seasonal fluctuations by, 321; New York bank clearings, 261, 268; use in correlating cycles in time series, 426; weighted, 270
Multiple and partial correlation, 485; limitations, 514; references, 515
Multiple correlation, application of the method, 499; coefficients of, 496; method valid for linear relationships, 499

National Industrial Conference Board, index numbers of the cost of living, 246
Nature, Uniformity in, 550
Net correlation, 500. *See also* Partial correlation
New York (State) Census of Manufactures, 406
New York State Department of Labor, index of weekly earnings, 249
Nitrogen as fertilizer, and wheat yield, 442, 443, 444
Non-linear correlation, 432; references, 452
Non-linear relationship, 20
Normal curve of error, 107, 108, 160, 516, 531; area in terms of abscissa, 538; economic application, 527; fitting, 532; illustration of the measurement of areas under, 539; ordinates, 532, 533; references, 546
Normal equations, 276; checks on the formation of, 573; derivation, 563; formation, 564; solution, 576; solution by the Doolittle method, 577, 578; standard set, 566
Normal values, monthly, 329, 330, 332
North Dakota, wheat yield, 114, 118
Numerical data on diagrams, 57, 58

Oats, production and price, 454, 455, 457, 467, 476, 477, 479, 480, 481, 483
Observations, 62; accuracy, 70; errors, 99; probable error, 160. *See also* Probable error
Ogburn, W. F., 269
Ogive, 91, 92; frequency curve and, relation, 93, 94

Ordinate, 12
Ordinates of the normal curve, 532, 533
Organization of data, 61

Parabola, 21; second degree, third degree, 25, 280, 281, 283
Parabolic curve, 20
Pareto, Vilfredo, law of, 140
Partial correlation, 500; computation of coefficient, 506, 507; distinguished from simple correlation, 502; method, 502
Peake, E. G., 105
Pearl, Raymond, 296, 297, 566
Pearson, Karl, 165, 166, 532, 537, 550; Chi-square test, 543; coefficient of correlation, 373; correlation ratio, 441, 450; formula for skewness, 166; frequency curves, 545, 546
Percentages on diagrams, 54
Percentiles, 124
Periodic fluctuations, 257
Periodic functions, 25
Perrin, Emily, 305
Personal incomes, 81, 121; distribution among recipients, 82; distribution of recipients, 83, 84, 85, 86. *See also* Taxable personal incomes
Persons, W. M., 192, 216, 218, 421; on advantages of method of link relatives, 328; commodity price index of business cycles, 238, 243; on the constitution of Bradstreet's index, 235; on correlation of time series, 419; index of trade for the United States, 361; index of wholesale prices, 223; measurement of seasonal fluctuations, 318; trend representation by related series, 306
Petroleum. *See* Coal and petroleum production
Petroleum production, 291; line of trend fitted to logarithms, 292, 293; fitted to natural numbers, 294, 295
Physical volume, index numbers, 344
Plotting paper, 33
Point, location of, 12
Population, curve of growth, 297
Potato prices, 475
Potential series, 25, 280
Preferred stock prices, 136
Prescott, R. B., 297
Price changes, 175; averaging, 187; wholesale commodity prices, 175
Price groups, 226
Price ratios, 179. *See also* Relative prices
Price system, 5, 8; elements, graphic representation, 224
Prices, 4; aggregates of actual, 193; Bradstreet's index numbers of wholesale, 173; harmonic mean in handling, 142; importance, 5; index numbers of (*see* Index numbers of prices); post-war trend, 306; variation, 148; weighted aggregates of actual, 207, 208. *See also* Relative prices
Probabilities, 9; *a priori* and empirical, 522; addition, 518; elementary theorems, 516; examples of simple, 517; measurement, 520; multiplication, 518
Probability curve, 107
Probable error, 160, 162, 164, 555, 556, 559
Product-moment method of linear correlation, 385, 390, 402
Production, 3, 4; absence of statistics, 344; actual and scheduled compared, 47, 48, 49, 50; costs, 46, 47; index numbers based on monthly values, 351; index numbers of E. E. Day, 347, 348, 349, 350, 351; index of War Industries Board, 345, 346; steel, 39
Products of factories in New York (State); number of wage earners as related to, value of, 406, 407
Profits, 5
Progressive mean, 270
Projections, 304

Quantitative data, 6, 7; classification, 97
Quantitative methods, 6, 8
Quartile deviation, 158, 161, 163
Quartiles, 124; graphic location, 131, 133, 134; symbols, 583

Random selection, 552
Range, 149, 161, 162

Rates, 135
Ratio chart, 33; advantages, 38; advantages, summary, 44
Ratios, 135, 137
Raw materials, 227; giving weight to various classes, 228
Reciprocals, for standard error and index of correlation, 468; in measurement of relationship, 465
Rectangular coördinates, 11; location of a point with reference to, 12
Reed, L. J., 297
References, list, 589
Regression equation, 394, 454; use, 399
Regression, lines of, 393, 394
Related series, 306
Relationship, measurement of, 363; between time series, 410; comparison, derived from arithmetic, geometric, and harmonic measures, 471; comparison of measures, 498; computation of measures of, 381; curves that may be used, 472; factors governing the choice of measures, 473; linear correlation, 363; multiple and partial correlation, 485; non-linear correlation, 433; problem of estimation and, 453; reciprocals, 465; symbols used, 465, 584
Relative prices, 179; arithmetic averages in construction of index numbers, 195, 196; distribution of, 222; commodities in 1900, 182, 183; distribution of, 346; commodities in 1914, 179, 180; distribution of, 1437 commodities in 1918, 185, 186; distribution of, 1437 commodities in 1918 plotted on logarithmic scale, 189; geometric averages, 198; harmonic averages, 199; medians for index numbers, 197; weighted arithmetic averages, 210; weighted geometric averages, 211, 212
Reliability, measures of, 553; limitations to measures of, 559
Research, 8
Residual, 274
Retail prices. *See* Index numbers of retail prices
Rietz, H. L., 153
Robertson, R. D., 433
Root-mean-square deviation, 154, 303, 368
Running, T. R., 301, 303

Sales, chart, 40, 41, 42; records, 7
Samples, 550
Sampling, 548; necessity of a representative sample, 552; references, 561; simple, conditions, 552; standard error of, 540, 541
Sawmills, 94, 95; earnings of band sawyers, 120
Scale for curves, reading, 56
Scales of a diagram, 57
Scatter, 109, 147, 372
Scatter diagrams, 364, 366, 404
Seasonal variations, 257; comparison of indices, 328; computing seasonal index numbers by averaging ratios to trend, 323; determining monthly trend values, 329, 330, 332; measurement, 315; method of link relatives, 318, 319; references, 343; use of moving averages, 321
Second degree parabola, 280, 281, 283; alfalfa and irrigation and, 434; business failures and, 286
Second order coefficients of correlation, 507; computation, 511
Secular trend, 256, 257; comparison of lines of trend, 288; measurement, 259; references, 313; relationship between trends, 411; representation by mathematical curves, 271; representation by related series, 306; selecting appropriate type, 300
Selected points, method of, 299
Semi-logarithmic charts, 32, 39. *See also* Ratio chart
Semi-logarithmic curves, 290
Series, 18; arithmetic, 18; continuous and discrete, 85; frequency, 61, 92; geometric, 23; potential, 25, 280; time, 37, 61. *See also* Time series
Sheppard, W. F., 158, 530, 537
Short term fluctuations, correlation, 427
Shots, distribution, 100, 101, 102
Simplification, 8
Sine curve, 20, 25, 26
Skewness, 110, 147; measures of, 166, 546; symbols, 583

Slide rule, 30
Smoothing of curves, 80, 87, 127
Snyder, Carl, 256
Soldiers, distribution by heights, 98
Spearman, Charles, 436
Speedwell Automobile Co., 47, 48
Standard deviation, 154, 162, 163, 368, 405, 513; computation, 154, 156
Standard error of estimate, 368, 375; application, 461; computation, 368, 372, 495; computation formula, 376; in logarithmic terms, 457; in terms of ratios, 460; index of correlation and, in terms of reciprocals, 468; interpretation, 459, 469; derivation of the formula for, 567
Standard error of sampling, 540, 541
Standard errors of chief statistical measures, 555
Statistical description, 548
Statistical induction, 548; references, 561
Statistical methods, 3; definition, 6; external problems and, 8; internal administration problems and, 6; limitations, 9
Statistical results, 549; generalization, 550
Statistical tables, 73. *See also* Tables
Steel production, 39
Steinmetz, C. P., 301; on empirical curves, 283, 284
Stewart, Ethelbert, 95
Stock fluctuations, 149
Stock price cycles as related to business cycles, 421, 423, 424, 425, 426
Straight line, 16; business failures and, 286; fitting, 273, 278
Symbols, glossary of, 582
Symmetry, 147

Tables, cumulative, 90; statistical, structure, 73
Taxable personal incomes in relation to motor vehicle registration, 364, 365, 366, 369
Telephone messages, 527, 528, 535, 536, 543
Telephone poles, 88, 89, 90, 91, 92, 130
Temperature and corn yield, 485, 486
Testing goodness of fit, 535, 540; Chi-square test, 543
Third degree parabola, 280, 282, 284; business failures and, 287
Thomas, Dorothy, 269
Time element, 256
Time rates, 141
Time reversal test, 203, 204, 212
Time series, 37, 61; analysis; measurement of seasonal and cyclical fluctuations, 315; analysis: measurement of trend, 252; data and preliminary organization, 253; difficulties in the correlation, 419; fluctuations, periodic and random, 257, 258; forces affecting, 255; general considerations relating to analysis, 342; graphic representation, 254; measurement of relationship, 410; references on correlation of, 431
Titles of diagrams, 59
Tolley, H. R., 491
Trade, index of volume, by Federal Reserve Bank of New York, 359, 361
Trend, 256; concept, 304, 306; measurement, 252; references, 313. *See also* Secular trend

Uniformity of nature, 550
U. S. Bureau of Labor Statistics, 207, 222, 253; index numbers of the cost of living, 246; index of retail food prices, 244; index of wholesale prices, 229, 243
U. S. Department of Agriculture, 246, 248; index numbers, 246, 247
U. S. Steel Corporation, unfilled tonnage, 254

Value, 8
Variability, measures of, 512
Variables, 14; correlation of several, 485; independent and dependent, 14; symbols for, 14, 582
Variation, 106; absolute and relative, 148; characteristic features of the chief measures, 162; coefficient of, 165; measure of, 109; measures of absolute, 148, 149; measurement of relative, 164; nature and significance, 147; references, 168; relations between different measures, 161; symbols, for measures of, 583
Visualization, 36

Volume of trade, Federal Reserve Bank index, 359, 361

Wage-earners, cumulative distribution, 133; number as related to value of their products, 406, 407
Wages, index numbers of money wages and real, 248; index of real, 250
Walsh, C. M., 140
War Industries Board, index numbers of wholesale prices, 233, 243; production index, 345, 346
Weekly earnings, 248; as basis of frequency distribution, 65, 72, 74; distribution, 132, 133, 134
Weighted aggregates of actual prices, 207, 208
Weighted arithmetic averages of relative prices, 210
Weighted arithmetic mean, 115
Weighted average, 115
Weighted geometric averages of relative prices, 211, 212
Weighting of index numbers, Fisher's four methods, 209
Weld, L. D., 527
Weldon, W. F. R., 523
Wheat flour exports, 38
Wheat, nitrogen fertilizer and yield, 442, 443, 444; No. 1 Northern Spring, relative price, index numbers, 170; yield in North Dakota, 114, 118
Whipple, G. C., 98
Wholesale prices, 306. *See also* Index numbers of wholesale prices
Working, Holbrook, 475

Yule, G. U., 71, 155, 446, 523, 552

Zero line, 52, 53
Zero order coefficient of correlation, 507
Zone of dispersion in artillery firing, 100, 101, 102
Zones of estimate, 384; significance, 477

311
N62

RAYMOND H. FOGLER LIBRARY
DATE DUE

	APR 1969		

UNIV OF MAINE
3 7550 01013627 4